The Dissertation Committee for Jessica Yingchich Guo certifies that this is the approved version of the following dissertation:

ADDRESSING SPATIAL COMPLEXITIES IN RESIDENTIAL LOCATION CHOICE MODELS

Committee:

Chandra R. Bhat, Supervisor

Stephen Donald

Randy B. Machemehl

Chandler Stolp

C. Michael Walton

ADDRESSING SPATIAL COMPLEXITIES IN RESIDENTIAL LOCATION CHOICE MODELS

by

Jessica Yinghchieh Guo, B.S., M.B.

Dissertation

Presented to the Faculty of the Graduate School of

the University of Texas at Austin

in Partial Fulfillment

of the requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December, 2004

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr Chandra Bhat, my supervisor, for his invaluable advice, guidance and encouragement. He has helped me to realize my potential as a researcher and to develop the necessary skills required to follow my dream of becoming an academic. It has been a privilege and pleasure to work with him. I extend my gratitude to all members of my dissertation committee for their constructive comments and suggestions during the course of my study. Many thanks must also go to Lisa Weyant for her kind assistance in administrative matters.

I am thankful for all my friends at UT, and those in Australia and Taiwan, for their support throughout the last four years. My deepest appreciation goes to my grandparents, Mom, Dad, Amber and Adam, who have been constantly supportive of my academic pursuit. I am also grateful to Mom and Dad Wilson and Gavin for their affection and good wishes. Last but not least, I thank my husband, Michael, who has shared with me the times of frustration and joy that accompanied my PhD study. This wouldn't have been possible without his love and support.

ADDRESSING SPATIAL COMPLEXITIES IN RESIDENTIAL LOCATION CHOICE MODELS

Publication No. _____

Jessica Yingchieh Guo, Ph.D.
The University of Texas at Austin, 2004

Supervisor: Chandra R. Bhat

Over the last two decades, there have been limited advances in the conceptualization of, and the modeling methodology for, the residential location choice problem. A widely used methodology for modeling individual household's residential choice is discrete choice analysis. Analysts typically consider administratively defined zones as discrete choice alternatives and apply the logit models to the residential choice problem in the same manner as for non-spatial contexts.

This research argues that there are distinctive features of the residential choice problem that distinguish it from non-spatial choice problems. Failure to account for these features may lead to erroneous analytical results and ineffective spatial policies. Two

important spatial features of the residential choice problem are addressed in this study. The first feature relates to the perceived similarity between neighboring choice alternatives that are intangible or difficult to quantify. To address the problem, this dissertation develops the mixed spatially correlated logit (MSCL) model by superimposing a mixing structure to accommodate unobserved heterogeneity across households over a closed form analytic structure that accommodates unobserved inter-alternative correlation. The empirical application of the model shows that the MSCL structure is both conceptually and statistically superior to the conventional modeling approach.

The second spatial issue addressed in this dissertation is the representation and measurement of spatial factors. By measuring spatial factors over administratively defined zones, the conventional grouped alternatives approach fails to relate the configuration of spatial units to decision makers' perception of space. The dissertation proposes a multi-scale structure to replace the conventional 'flat' approach. The proposed structure is innovative in that it allows the choice factors' spatial extent of influence be determined endogenously. In addition, the multi-scale model can be used to test alternative hypothetical representations of neighborhoods as perceived by different households for different residential alternatives. The empirical application of the model demonstrates that social-economic and demographic factors generally have a smaller spatial extent of influence on residential choice than land-use factors. The results also show differing effects of choice factors when different spatial definitions are employed, suggesting the need for future research on behaviorally-realistic spatial representations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

The home is where people typically spend most of their time, a common venue for social contact and, for most people, a major financial and personal investment. One's choice of residence also reflects one's choice of the surrounding neighborhood, which has a significant impact on one's well-being and quality of life. The topic of residential location choice has, therefore, been of interest to sociologists, psychologists, urban economists, geographers and transportation planners. There is a substantial body of literature on the subject, covering both theoretical and empirical investigations from different perspectives, including the relationship between life quality and location, market differentiation in housing demand, societal value of urban amenities and neighborhood quality, and effects of spatial policies. Residential location preference as a factor in urban development patterns is important especially because of the high rate of residential mobility in the United States. According to the Census Bureau, between 1970 and 1999, every year there is an average of 17% of the national population who change their place of residence. Since then, this rate has consistently stayed above 14%. Between 2002 and 2003, approximately 83% of the residential moves in the country were within the same state and 63% were within the same county. These statistics suggest that, at least in terms of consumer demand, there is a strong potential for rapid shifts in housing patterns and perhaps equal potential for slowing or even reversing the trend (Weisbrod *et al*, 1980).

For urban and transportation planning, the concern for the causes and consequences of individuals' choice of residence arises from the recognition that it is the values, decisions and

actions of the people who are attracted to certain types of land use patterns that ultimately shape the transportation, land-use and urban form. The decision of residential location not only determines the connection between the household with the rest of the urban environment, but also influences the household's activity time budgets and perceived well being. Altering land use characteristics by itself might not affect the residents' travel behavior as expected by proponents of New Urbanism. Rather, travel characteristics might only change after new residents are attracted by new land use and move into an area while old residents who find the land use unsuitable eventually move out (Kitamura, Mokhtarlan and Laidet, 1997; Krizek and Waddell, 2002; Bagley and Mokhtarian, 2002; Lund, 2003). The need for understanding land use – transport linkage at the individual level and the debate over whether the influence of urban form is entirely due to individuals placing themselves into residential neighborhoods that support their travel propensities points to the need for better models of residential location preferences.

Over the past four decades, there has been considerable development in the mathematical modeling of residential activities. The earliest theoretical work on urban housing markets established a tradition of viewing housing in terms of a dwelling's market value or rent. The pattern of residential location in a city is explained by the trade-off between transport costs (which increases with distance from the city center) and housing costs (which decreases with distance from the city center) (Hoover and Vernon, 1959). Based on the 'trade-off' theory, Alonso (1964) was the first one to consider residential location choice based on the concept of utility maximization. The level of utility a household experiences depends on the expenditure in goods, size of the land lots, and distance from the city center. Aloson's model was later extended by economists such as Muth (1969), Mills (1972), Evans (1973) and Wheaton (1974) and these models are referred as the classical urban land market models. The most criticized aspects of these urban economic studies are: (1) The models treat location as a one-dimensional variable

(i.e., distance from major employment centers) and are therefore incapable of handling the common situations of dispersed employment centers and asymmetric development patterns (Waddell, 1996); (2) All members of any one socio-demographic class are considered to have identical behavior, which is certainly an oversimplification of reality. (3) By reducing the complexity of the housing commodity, which is multidimensional and heterogeneous, to the one-dimensional measure of price, one assumes that many of the important and interesting housing market phenomenon are irrelevant (Kain and Quigley, 1975).

The desire to identify the extent to which various dimensions of housing affect the housing price has spawned innumerable applications of the hedonic pricing model, originally developed by Griliches (1961) and Rosen (1974). The approach involves regressing the housing payments for a residential site on a vector of housing attributes, including both the physical characteristics of the dwelling (size, appearance, features) and the surrounding neighborhood (accessibility to schools and shopping, quality of other houses, availability of public services). The parameter estimates indicate the implicit prices possessed by the housing attributes. The primary limitation of the hedonic approach is that the analysis is based on housing units, not the households. As a result, no household-level commute or other household characteristics are present to account for variations among households' preferences for various dwelling and neighborhood attributes. This limitation makes the hedonic pricing approach inappropriate for understanding the land-use transport interaction at the individual decision maker's level.

While the urban economic and the hedonic pricing models are intended to provide an economic explanation for residential patterns, the spatial interaction models are intended not so much as to provide a theoretical basis, but to serve as operational tools for predicting spatial patterns. The spatial interaction modeling approach does so by drawing an analogy between the interaction of human activities in space and the Law of Gravity in Physics. The approach was

first applied to predict the location of population in residential zones in an urban region by Hansen (1959), who assumed that the accessibility to employment is the principle determinant of the location of population. The landmark model of this approach is the Lowry model (Lowry, 1964) developed for the Pittsburgh metropolitan area that involves iteratively allocating residential population and employment to the zones in the study region until the constraints on land use are met and the distribution of population and employment reflects the resulting travel between zones. Variants of Lowry's model have served as core land-use forecasting models in a large number of aggregate land-use and transportation studies. Despite its practical popularity, the spatial interaction model has received heavy criticisms for simply representing and reproducing empirical regularities and not providing a theoretical explanation of the factors, in addition to accessibility, that account for spatial interactions (Romanos, 1976; Sayer, 1976; Briassoulis, 2000).

The aforementioned approaches to studying households' residential choice are either too aggregate in nature or lacking in behavioral foundation. The discrete choice analysis approach introduced by McFadden (1974) avoids both of these shortcomings. The approach is motivated by the desire to understand the behavioral process that leads to an agent's choice among a set of options. A decision maker is assumed to consider various factors that collectively determine the utility obtainable from each alternative and chooses one that yields the maximum utility – a behavioral assumption similar to that underlying the urban land market models. The pioneering work of Lerman (1975) and McFadden (1978) lead to the extensive use of discrete choice analysis in studying residential choice. The popularity of this approach is attributed to at least the following two reasons. First, the discrete choice models provide a way of understanding, at the disaggregate level, how a household trades-off among the wide range of choice factors that come into play. Second, the approach allows the analyst to examine the choice behavior based on both

accepted and rejected alternatives and to relate spatial behavior to locational characteristics as well as the complex attitudes, preferences and tastes of individuals. The modeling results can thus help devise urban policies that effectively target specific population groups. For these reasons, discrete choice analysis dominates spatial choice theory (Thill and Wheeler, 2000), even though it was originally developed for non-spatial contexts such as the choice of transportation mode. In addition to its application to residential location choice, the approach has been employed to model consumer store choice (Fotheringham, 1988; Rust and Donthu, 1995), tourist destination (Eymann and Ronning, 1997), recreational demand (Feather, 1994; Train, 1998; Parsons and Hauber, 1998; Pozsgay and Bhat, 2002), industrial location (Hansen, 1987), criminals' site selection (Xue and Brown, 2003), and interregional migration (Kanaroglou and Ferguson, 1998; Pellegrini and Fotheringham, 2002).

1.2 Features of Spatial Choice Problems

When analyzing residential location or other types of spatial choice behavior, past modeling efforts typically apply the discrete choice models in the same manner as they would do for a non-spatial context with little modification (Pellegrini and Fotheringham, 2002). However, spatial choice contexts exhibit several distinct features not found typically in their non-spatial counterparts that can create problems for the use of standard discrete choice models. These characteristics are summarized below:

1. *Definition of alternatives*: Contrary to most aspatial contexts, spatial choice problems often involve choice elements that are difficult to define (Lerman, 1983; Fischer and Nijkamp, 1987). For example, when a person chooses where to shop, is she/he selecting a specific store, a neighborhood populated with shops, or a specific shopping mall? Similarly, tourists choosing a holiday destination may be selecting among one or more

different geographical levels, such as a hotel, a city, or a country. The definition of the choice set is far from trivial for such spatial applications.

2. *Definition of choice set*: In spatial choice situations, decision makers often face a very large set of potential options. However, in practice, the number of alternatives actually considered is constrained by the individuals' limited capacity for gathering and processing information (Fotheringham et al, 2000). According to Bettman (1979), this limit might be reached with as few as six or seven alternatives. Consequently, it would be fairly unrealistic to assume individuals can evaluate all possible alternatives at any one time. The identification of individual choice sets is therefore a challenge to the analyst (Kanaroglou and Ferguson, 1998).
3. *Substitutability among choice alternatives*: Due to the continuity of space, the spatial alternatives faced by decision makers are likely to follow the First Law of Geography postulated by Tobler (1970), that everything is related to everything else, but closer things are more closely related. An alternative at a given location may be perceived as more similar, and therefore more substitutable, to an alternative closer by rather than farther away. The perceived similarity between neighboring spatial alternatives are often intangible or difficult to quantify. Failure to account for such perceived similarity would lead to inaccurate interpretations of choice behavior. Yet, in standard discrete choice models, accommodating unobserved similarity among the choice alternatives is not a straightforward task.
4. *Measurement of spatial variables*: As in the case of other modeling efforts, the success of a discrete choice modeling exercise relies on correct model specifications, which are tied closely to accurate representation, or measurement, of relevant variables. For variables that are spatial in nature (which is the case for spatial choice problems and

also not uncommon for non-spatial choice problems), their value can be observed only after a location has been specified (for point variables) or a space been demarcated (for areal variables). In the latter case, the continuity of space renders almost infinitely many ways for an analyst to define areal units for measuring. Without knowing which of the many spatial configurations to use, past efforts of spatial choice modeling typically use administratively defined spatial units, such as census tracts, for which data are readily available. These administrative units often bear no relation to how the decision makers themselves measure, or perceive, the spatial factors in their mind. Such a practice may easily lead to inaccurate analytic outcomes.

Because of the issues identified above that set the spatial choice problems apart from the non-spatial ones, 'discrete spatial choice analysis' is becoming a distinct area of research beyond the umbrella of its origins in non-spatial discrete choice modeling (Pellegrini and Fotheringham, 2002). Any analysis that hopes to provide a robust explanation for residential location choice dynamics and a framework for evaluating housing policy must take these distinguishing features seriously.

1.3 Research Objectives

The general goal of this dissertation is to address the added complexities introduced by space in the discrete choice analysis of residential location. The research focus is on two of the issues identified in the preceding section: the substitutability among choice alternatives and the measurement of spatial factors.

The issue of substitutability is rooted in the Independence from Irrelevant Alternatives (IIA) assumption often embedded in residential location choice models. The assumption leaves

no room for accommodating any perceived similarity shared by the alternative locations that is unobserved by the analyst. The first research objective of this dissertation is thus to treat the IIA problem by drawing on and integrating advanced modeling techniques that allow for more flexible correlation structures among choice alternatives.

The second research objective is to clarify what decision makers mean by 'location' when choosing a residence and so to develop ways of appropriately measuring and incorporating location factors in the choice model. Basic to this research is the belief that analysts should measure what matters to people over the area that really matters to people. Only when the choice factor is measured over its true extent of influence can its exact effect on residential location choice be econometrically extracted.

1.4 Dissertation Outline

This dissertation is organized as follows. Chapter 2 discusses the standard discrete choice modeling framework used for residential location choice analysis and previous empirical studies based on such a framework. The survey is intended to provide an understanding of the state of the art. The major findings derived from these studies and the choice determinants found to affect residence choice behavior are also presented.

Based on observations drawn from the literature survey in Chapter 2, Chapter 3 describes in detail the causes and origins of the two spatial complexities which the dissertation aims to address. The chapter further discusses how existing conceptual and methodological frameworks may, or may not, be used to resolve these complexities.

Chapter 4 develops the proposed methodology for accommodating unobserved similarity shared among neighboring residential locations. The structure, property and estimation of the

proposed model are described. An empirical application of this model and the implications of the results are also discussed.

Chapter 5 describes a hierarchical modeling structure proposed for treating the spatial measurement problem. The chapter discusses the different empirical results obtained from application of the conventional discrete choice model and the proposed model. It also presents a comparative analysis of empirical results derived from using two different sets of spatial units to measure location factors.

Chapter 6 summarizes the main findings and addresses the limitations of the proposed methodologies. It concludes by outlining a number of directions in which the proposed methodologies could be further extended.

CHAPTER 2

DISCRETE CHOICE MODELS OF RESIDENTIAL LOCATION

Urban housing units differ profoundly over a great many dimensions that are essential elements of consumer choice. Housing units differ from each other structurally with respect to condition, architectural style and features, size, plumbing facilities, and the like. They also differ in their accessibility to other relevant parts of the urban area; in size, shape, topography, and so forth of the lot of land on which a unit sits; and in the socio-demographic character of, and level of public services provided in, their neighborhoods. Different households have different tastes for the various dimensions of the housing package, and typically each household conducts an extensive search for a unit appropriate to its taste and income.

The nature of the residential choice problem described above makes the discrete choice modeling approach an appropriate tool for investigating how households of various types distribute themselves over different housing types and over space in response to market forces and policies. This chapter provides an understanding of the state of the art in discrete choice modeling of residential location choice. Section 2.1 describes the random utility maximization (RUM) framework from which discrete choice models are derived. Section 2.2 presents the model structures used extensively in past studies of residential choice. Methods for estimating these models are explained in Sections 2.3 and 2.4. Statistical tests for hypothesis testing and for comparing model goodness of fit are described in Section 2.5. Section 2.6 reviews past modeling efforts of residential choice. The commonalities and differences among these efforts are discussed and their findings are summarized.

2.1 Random Utility Maximization Framework

The origin of the RUM framework goes back to Thurstone (1927), who developed the Law of Comparative Judgment. In his theory, the perceived level of a psychological stimulus equals its objective level plus a random error. The probability that one object is judged higher than the other is the probability that this object has the higher perceived stimulus. When the perceived stimuli are interpreted as levels of satisfaction, or utility, Thurstone's theory can be interpreted as a model for economic choice in which utility levels are random, and the observed choice is the alternative that has the highest realized utility level. This connection was made in by Marschak (1960), who called this the random utility maximization hypothesis.

The RUM principle is described formally as follows. Denoting the utility that decision maker n obtains from alternative j by $U_{n,j}$, $j = 1\ldots, J$, the decision maker chooses alternative i if and only if $U_{n,i} > U_{n,j}, \forall j \neq i$. Since the analyst observes only some of the factors considered by the decision maker, the utility is modeled as comprising a deterministic component $V_{n,j}$ and a stochastic component $\varepsilon_{n,j}$:

$$U_{n,j} = V_{n,j} + \varepsilon_{n,j}. \qquad (2.1)$$

The deterministic component, $V_{n,j}$, is typically specified as a function of the observed attributes of alternative j as faced by decision maker n, denoted as $X_{n,j}$, and unknown parameters β that are to be statistically estimated. The stochastic component $\varepsilon_{n,j}$ captures the factors unobserved by the analyst and is therefore treated as random. By assuming that the random vector $\varepsilon_n = \left(\varepsilon_{n,1}, \varepsilon_{n,2}, \cdots \varepsilon_{n,J}\right)$ follows a joint probability density function $f\left(\varepsilon_n\right)$, the probability that decision maker n chooses alternative i can be stated as:

$$
\begin{aligned}
P_{n,i} &= \text{Prob}\left(U_{n,i} > U_{n,j}, \forall j \neq i\right) \\
&= \text{Prob}\left(V_{n,i} + \varepsilon_{n,i} > V_{n,j} + \varepsilon_{n,j}, \forall j \neq i\right) \\
&= \text{Prob}\left(\varepsilon_{n,j} - \varepsilon_{n,i} < V_{n,i} + V_{n,j}, \forall j \neq i\right) \\
&= \int_{\varepsilon_n} I\left(\varepsilon_{n,j} - \varepsilon_{n,i} < V_{n,i} + V_{n,j}, \forall j \neq i\right) f\left(\varepsilon_n\right) d\varepsilon_n, \qquad (2.2)
\end{aligned}
$$

where $I(\cdot)$ is the indicator function taking the value of 1 if the condition in parentheses is true, and 0 otherwise. Different specification of the density function $f\left(\varepsilon_n\right)$ derived from different assumptions about the distribution of the unobserved component of utility lead to different choice models.

2.2 Conventional Model Structures

2.2.1 Multinomial Logit Model

The conditional or multinomial logit model (MNL), originally developed by McFadden (1974), is by far the most widely used discrete choice model derived from the RUM principle. It is derived under the assumption that each stochastic term $\varepsilon_{n,j}$ is independently and identically distributed (IID) with a type I extreme value (or Gumbel) distribution (Johnson and Kotz, 1970), described by the following density function:

$$
f\left(\varepsilon_{n,j}\right) = e^{-\varepsilon_{n,j}} e^{-e^{-\varepsilon_{n,j}}} \qquad (2.3)
$$

Substituting the density function in Equation (2.3) into Equation (2.2), one arrives at a closed form expression for the logit choice probability:

$$
P_{n,i} = \frac{e^{V_{n,i}}}{\sum_j e^{V_{n,j}}} \qquad (2.4)
$$

If $V_{n,j}$ is linear in parameters, the choice probability becomes:

$$P_{n,i} = \frac{e^{\beta' X_{n,i}}}{\sum_j e^{\beta' X_{n,j}}}. \tag{2.5}$$

The direct elasticity, *i.e.* the percentage change in the choice probability of alternative i due to a 1% change in the q th variable associated with alternative i, is:

$$\frac{\partial P_{n,i}}{\partial x_{n,iq}} \cdot \frac{x_{n,iq}}{P_{n,i}} = \left(1 - P_{n,i}\right)\beta_q x_{n,iq}. \tag{2.6}$$

Similarly, the cross elasticity, *i.e.* the percentage change in the choice probability of alternative i due to a 1% change in the q th variable associated with alternative j, is:

$$\frac{\partial P_{n,i}}{\partial x_{n,jq}} \cdot \frac{x_{n,jq}}{P_{n,i}} = P_{n,j}\beta_q x_{n,jq}. \tag{2.7}$$

The independence assumption about the stochastic terms implies that the unobserved component of utility for one alternative is unrelated to the unobserved component of utility for another alternative. This is in accordance with Luce's (1959) choice axiom of independence from irrelevant alternatives (IIA), which states that the relative probabilities of any two alternatives depend only on their relative utilities and are independent of other alternatives of the choice set. As shown in Equation (2.7), the IIA property also implies that an improvement in one alternative draws proportionately from all the other alternatives. This phenomenon is referred as proportionate substitution. An alternative can thus be introduced into or eliminated from the choice set without the re-estimation of the utility function parameters and the choice probabilities. Proportionate substitution greatly facilitates forecasting and is an attractive feature for many discrete choice problems. However, in cases where there are alternatives that are close substitutes for each other, such as the well-known 'red bus – blue bus' problem, the IIA assumption becomes

unrealistic, resulting in inconsistent estimates of the model parameters and of the choice probabilities (Horowitz, 1981).

The MNL model is also limited in its power to represent preferential variations among decision makers. In general, the value that decision makers place on each attribute of the alternatives varies across decision makers. This taste variation can be explicitly captured in the logit models by allowing in the observed utility, $V_{n,j}$, the observed attributes of alternatives, $X_{n,j}$, to interact with the observed attributes of decision makers, $S_{n,j}$. However, when there is taste variation attributed to decision makers' individual attitudes and experiences that are unobserved by the analyst, treating the β parameters as constant across the population would be a misspecification. Not only the MNL model does not provide estimates about the distribution of tastes, there is also no guarantee that the β estimates will provide adequate approximations of the average tastes of the population (Chamberlain, 1980; Train, 2003).

2.2.2 Nested Logit Model

Recognizing that the unobserved similarities among the alternatives invalidate the MNL model, McFadden (1978) proposed the nested logit (NL) model, which is also RUM-based, to model residential location choice. His conceptual framework considers groupings of dwellings that share similar unobserved characteristics as *communities*. The utility assigned by decision maker n to dwelling j in community c, $c = 1\ldots,C$, alternative (c, j) is denoted as:

$$U_{n,cj} = V_{n,cj} + \varepsilon_{n,cj} . \tag{2.8}$$

The NL model is obtained by assuming that the vector of unobserved utility, $\varepsilon_n = \left(\varepsilon_{n,1}, \varepsilon_{n,2}, \cdots \varepsilon_{n,J}\right)$, has cumulative distribution:

$$f\left(\varepsilon_n\right) = \exp\left(- \sum_{c=1}^{C} \left(\sum_{j \text{ in } c} e^{-\varepsilon_{n,j}/(1-\sigma_c)} \right)^{1-\sigma_c} \right). \tag{2.9}$$

This distribution is a generalization of the distribution described by Equation (2.3) that gives rise to the MNL model. For the MNL model, each stochastic term, $\varepsilon_{n,j}$, is independently univariate extreme value distributed. In Equation (2.9), the marginal distribution of each $\varepsilon_{n,j}$ in the NL model is univariate extreme value, with the $\varepsilon_{n,j}$'s correlated for dwellings in the same community but not across communities. The parameter σ_c ($0 \le \sigma_c < 1$) measures the level of correlation among dwellings nested under community c, with $\sigma_c = 0$ indicating complete independence within the community.

Suppose $X_{n,cj}$ and $Y_{n,c}$ are vectors of observed dwelling- and community-specific attributes, respectively, and $V_{n,cj}$ assumes an additively separable, linear-in-parameters form:

$$V_{n,cj} = \beta' X_{n,cj} + \alpha' Y_{n,c} . \tag{2.10}$$

The choice probability of alternative (c, j) given rise from Equation (2.9) is:

$$P_{n,ci} = \frac{e^{\left(\beta' X_{n,ci} + \alpha' Y_{n,c}\right)/(1-\sigma_c)} \left(\sum_{j \text{ in } c} e^{\left(\beta' X_{n,cj} + \alpha' Y_{n,c}\right)/(1-\sigma_c)} \right)^{-\sigma_c}}{\sum_{b=1}^{C} \left(\sum_{j \text{ in } b} e^{\left(\beta' X_{n,bj} + \alpha' Y_{n,b}\right)/(1-\sigma_b)} \right)^{1-\sigma_b}} . \tag{2.11}$$

For the ease of interpretation, one can consider the above expression as the product of a marginal probability $P_{n,c}$ and a conditional probability $P_{n,i|c}$ that both take the form of logits:

$$P_{n,c} = \frac{e^{\alpha' Y_{n,c} + (1-\sigma_c) I_c}}{\sum_{b=1}^{C} e^{\alpha' Y_{n,b} + (1-\sigma_b) I_b}} , \tag{2.12}$$

$$P_{n,i|c} = \frac{e^{\beta' X_{n,ci}/(1-\sigma_c)}}{e^{I_c}} , \tag{2.13}$$

where

$$I_c = \ln \sum_{j=1}^{J_c} e^{\beta' X_{n,cj}/(1-\sigma_c)} . \quad (2.14)$$

The quantity I_c is referred as the inclusive value and represents the expected utility that decision maker n receives from choosing among the dwellings in community c (Train, 2003).

2.2.3 Grouped Alternatives Model

In an analysis of mobility bundle choices, Lerman (1975) argues that, in practice, data about individual housing units are typically not available and most surveys provide data only about which community a household selected rather than which specific housing unit. Thus, he proposes a grouped alternatives choice model of the form:

$$P_{n,c} = \frac{e^{\alpha' Y_{n,c} + \beta' \overline{X}_{n,c} + (1-\sigma_c)\ln J_c}}{\sum_{b=1}^{C} e^{\alpha' Y_{n,b} + \beta' \overline{X}_{n,b} + (1-\sigma_b)\ln J_b}}, \quad (2.15)$$

where $Y_{n,c}$ is a vector of community-specific attributes as defined before and $\overline{X}_{n,c}$ is a vector of the expected values of the observed dwelling attributes in community c:

$$\overline{X}_{n,c} = \frac{1}{J_c} \sum_{j=1}^{J_c} X_{n,cj} . \quad (2.16)$$

Lerman's model can be viewed as a MNL model where the choice alternatives are the groups of housing units, rather than the individual units. The term J_c in Equation (2.15) is used to correct for the grouping process such that, other conditions being equal, a large grouping would have a higher probability of being selected than a small grouping. It is thus often referred as the size variable. According to Lerman (1975), the grouping can be defined based on the dwellings' physical proximity, as well as along other dimensions of their measurable attributes. For example, instead of using census tracts (as proxy for communities) as the unit of grouped locations, one might consider levels of neighborhood quality such as a three-level status classification of upper, middle and lower classes. Alternatively, one might group the alternative

dwellings based on their structure type, such as single-family detached units, multiplex units and apartments, to reflect the sectoral housing submarket.

Lerman's model can also be viewed as a special case of McFadden's NL model. Specifically, in Equation (2.15), if the community c is homogeneous in terms of the attributes of the constituting dwellings so that $X_{n,cj} = \overline{X}_{n,c}, \forall j = 1 \ldots J_c$, the NL model defined by Equations (2.12) to (2.14) is exactly Lerman's group alternatives model. This link establishes the consistency of the group alternatives model with RUM under the homogeneity condition. However, if the homogeneity condition is not satisfied, which is often the case in practice, the parameter estimates will be biased and the degree of estimation error depends on the variances of $X_{n,cj}$. Therefore, in most cases, the grouped alternatives model is only an approximation of the 'theoretically correct choice model' (Quigley, 1985), where every distinguishable dwelling is treated as a distinct choice entity.

2.3 Estimation Techniques

Since the choice probabilities given by the logit models described in the previous section take a closed form, the estimation of these models can be pursued using the maximum likelihood (ML) method. The probability of decision maker n choosing the alternative that was in fact chosen can be expressed as:

$$\prod_{j=1}^{J} (P_{n,j})^{\delta_{n,j}}, \tag{2.17}$$

where

$$\delta_{n,j} = \begin{cases} 1 \text{ if decision maker } n \text{ chose dwelling (or dwelling group) } j \\ 0 \text{ otherwise} \end{cases}.$$

Assuming that each decision maker's choice is independent of others', the joint probability of each of the N decision makers in the sample choosing the alternative that was actually chosen is:

$$L(\theta) = \prod_{n=1}^{N} \prod_{j=1}^{J} (P_{n,j})^{\delta_{n,j}} . \tag{2.18}$$

The log-likelihood function is then:

$$LL(\theta) = \sum_{n=1}^{N} \sum_{j=1}^{J} \delta_{n,j} \ln P_{n,j}$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{J} \delta_{n,j} \ln \frac{e^{V_{n,i}}}{\sum_{j} e^{V_{n,j}}} . \tag{2.19}$$

In Equations (2.18) and (2.19), θ denotes the vector containing all the parameters in the model. The value of θ that maximizes the log-likelihood function is the ML estimator.

An alternative estimation procedure for the logit models is the maximum score (MS) method developed by Manski (1975). The MS estimator is one that satisfies the following condition:

$$\max_{\theta} = \sum_{n=1}^{N} \sum_{j=1}^{J} \delta_{n,j} , \tag{2.20}$$

where

$$\delta_{n,j} = \begin{cases} 1 & \text{if } i \text{ is chosen and } V_j \geq V_i, \forall i \neq j \\ 0 & \text{otherwise} \end{cases} . \tag{2.21}$$

In non-mathematical terms, the estimator is one that maximizes the number of times the alternative with the greatest utility was selected. The advantage of the MS method over the ML method is that the former does not require any specific distribution assumptions made about the stochastic utility, as long as the alternative with the highest utility also has the greatest probability of being selected. The drawback is that the MS estimators are neither asymptotically efficient nor

normal. Thus, it is impossible to perform asymptotic statistical tests of the significance of coefficients or of linear restrictions on the parameters (Lerman, 1975). Therefore, the ML method remains the most popular procedure for estimating discrete choice models.

2.4 Sampling of Alternatives

One of the features of spatial choice problems, as stated in Section 1.2, is that decision makers often face a very large set of alternative locations. This renders the estimation of model parameters a very expensive or even impossible task (McFadden, 1978; Train, 2000; Nerella and Bhat, 2003). Fortunately, with a logit model, estimation can be performed on a subset of alternatives without inducing inconsistency (McFadden, 1978). The "estimation by sampling of alternatives" procedure is described below.

Denote $f(D_n|i)$ as the sampling rule for obtaining a subset of alternatives D_n, conditional upon the observed choice of dwelling unit i, for decision maker n. This sampling rule needs to satisfy the property that, if $f(D_n|i)>0$, then $f(D_n|j)>0$. In other words, if a rejected alternative j is assigned to the subset D_n, then it is logically possible that j could have been the observed choice. Then, the maximization of the modified likelihood function,

$$LL(\theta)=\sum_{n=1}^{N}\sum_{j\in D_n}\delta_{n,j}\ln\left(\frac{e^{V(i)+\ln f(D_n|i)}}{\sum_{j\in D_n}e^{V(j)+\ln f(D_n|j)}}\right), \tag{2.22}$$

yields a consistent estimator for the logit model. However, since information is lost about alternatives not included in D_n, the estimators are not efficient.

The most common sampling approach is based on the condition that $f(D_n|i)$ is the same for all $j\in D_n$. This "uniform conditioning property" (McFadden, 1978) occurs if, for example, all non-chosen alternatives are assigned with an equal probability of being selected into D_n, so that the probability of selecting j into D_n when i is chosen by the decision maker is the same as the probability of selecting i into D_n when j is chosen (Train, 2000). When this property holds, the $\ln f(D_n|j)$ terms disappear from Equation (2.22) and the likelihood function to

maximize becomes the same as the one given in Equation (2.19) except that the subset of alternatives D_n replaces the complete choice set.

2.5 Statistical Tests

Once a model is estimated using the ML procedure, its goodness of fit can be measured by the likelihood ratio index, which is defined as:

$$\rho = 1 - \frac{LL(\hat{\theta})}{LL(0)}, \tag{2.23}$$

where $LL(\hat{\theta})$ and $LL(0)$ are the values of the log-likelihood function at the estimated parameters and at zero, respectively. The value of σ ranges from zero, when the estimated model does no better than no model at all (zero parameters), to one, when the estimated model provides a perfect prediction for all sampled observations. Any value between zero and one, however, has no intuitively interpretable meaning (Train, 2000). Thus, the likelihood ratio index is not at all similar in its interpretation to the R^2 used in regression. When comparing two model specifications based on the same sample and identical set of alternatives, one can compare the goodness of fit of the models using the likelihood ratio index. In the case when the number of parameters used in the models is different, the adjusted likelihood ratio index (Ben-Akiva and Lerman, 1985) can be used instead:

$$\rho^* = 1 - \frac{LL(\hat{\theta}) - K}{LL(0)}, \tag{2.24}$$

where K is the number of parameters in the estimated model.

As for testing hypothesis about individual parameters in discrete choice models, one can use the standard t-statistics as with linear regressions. A parameter estimate is significant at, say, 5% level (less than 5% chance that the associated difference from zero is due to random effects) when the corresponding t-statistic has an absolute value greater than 1.96.

2.6 Previous Modeling Efforts

There is an abundance of studies that attempt to understand the residential choice behavior through discrete choice models. As summarized in Table 2.1, these studies are common and/or different in the choice dimensions modeled, the model structure utilized, the way in which alternative residential location choices are represented, the study region and the population segment examined, the choice determinants considered, and the findings concluded from the empirical work. These commonalities and differences are discussed in the subsequent sections. Specifically, Section 2.6.1 identifies the choice dimensions considered and the model structures utilized in these past studies. The way in which choice alternatives are defined and the measurement of spatial factors describing the alternatives are discussed in Sections 2.6.2 and 2.6.3, respectively. Section 2.6.4 provides a list of the choice factors considered in the previous studies. The statistical significance and the behavioral interpretations of these factors are discusses. The discussion serves as a guide to utility specification for the empirical analyses performed in this study.

Table 2.1 Previous studies of residential location choice

Study	Study region and sample	Dimensions modeled	Model Structure	Residential Choice alternatives	Housing (H) and neighborhood (N) measures	Main findings
Abraham and Hunt (1997)	Calgary 961 households	Residential location, work location, and commute mode choice	NL	Zones (exact definition not specified)	• average housing price (N) • portion of sales that were not condominiums (N) • size variable (N) • commute time (N)	• commute time has 1.65 times the effect on household utility for women as for men • the average household is willing to pay an extra $10.44 per square meter of housing to be 1 min closer by car to a 40-year-old man's workplace, and an extra $17.38 per square meter to be 1 min closer by car to a 40-year-old woman's workplace
Anas (1981), Anas and Chu (1984)	Chicago	Residential location choice and mode choice	MNL and NL	Quartersetions (0.5X0.5-mile square zones)	• zonal average rent per dwelling (N) • zonal average housing age (N) • distance to the CBD (N) • quality of locatioanl and housing amenities (N)	• values of housing rent, travel time and travel cost elasticities are sensitive to model estimation methods • ranges of elasticity values are in agreement with those estimated for other cities • estimation uncertainty is comparable to that due to model specification
Ben-Akiva and Bowman (1998)	Boston 1259 households	Residential choice and activity-scheduling decisions	NL	TAZ	• expected utility of household members' daily activity schedule (H) • composite impedance measure for commute (N) • violent crime rate (N) • residential density (N) • income remaining after housing expenses (N) • school education performance (N) • proximity to industrial acreage (N) • town's expenditure on culture and recreation (N) • residential tax rate (N) • geographic indicator (N) • size variable (N)	• worker's accessibility has strong and positive effect on residential choice • the composite impedance for commute explains residential choice behavior better than the expected utility of activity schedule • households are less likely to reside in locations with high crime rate, high density, and high housing price relative to their income level • school performance, industrial acreage, expenditure on culture and recreation, tax rate, and cbd indicator are not significant choice determinants
Boehm (1982)	Michigan 1864 households	Tenure, dwelling size, neighborhood quality	NL	High- vs. low-income neighborhoods	• dwelling size (H) • average income (N) • average price of owned housing (N) • ratio of cost of owning versus cost of renting (N) • ratio of expected rate of change of housing prices to all prices (N) • price of a large unit relative to a small unit (N) • price of a unit in high income neighborhood relative to one in a low income neighborhood (N) note: neighborhood characteristics	• a larger family size increases the probability of choosing a large unit • a large family is less likely to live in high-income neighborhoods • the relative price of high versus low quality housing has no significant impact on the choice of neighborhood quality • blacks are less likely to live in high-income neighborhoods • an increase in household income can have a negative effect in certain homeowners and a positive effect on certain renters

					measured over census tracts	
Chattopadhyay (2000)	Chicago 3044 homes from 659 census tracts located in 103 city districts	Residential choice	3-level NL (dwelling, , neighborhood, city)	Dwelling, non-spatial groupings of neighborhood types (4), and city types (6 or 8)	• housing price (H) • #rooms (H) • age of dwelling (H) • area of lot (H) • presence of central air-conditioning (H) • parking facility (H) • racial composition (N) • median income (N) • distance from Chicago loop (N) • pollution (N) • property tax rate (N) • per-capita municipal spending (N) • per-capita spending by school districts (N) • geographic constants (N)	• whites opt for less #rooms, older houses, bigger lot size than non-whites • whites prefer living in a neighborhood with more white population, higher median income, farther from the CBD, and at a location with less air pollution • whites as well as large families prefer more for a city with lesser property taxes, better public schools and less municipal spending • large families prefer more rooms, older houses, and bigger lots than small families • large families opt for a neighborhood with higher percentage of white population, lower median income, farther from the CBD, and less pollution • measures of benefits of changes in neighborhood attributes are very sensitive to groupings of neighborhoods into types
Clark and Onaka (1985)	San Joseph county, South Bend, IN 289 renter households	Residential mobility, neighborhood, dwelling type, and dwelling unit	NL	Residential sectors defined by the analysts	• log of mean zonal household income divided by gross income of a household (N) • racial composition (N) • %dwelling units built between 1960 and 1970 (N) • log of commute distance (N) • expected out-of-pocket moving cost (H) • expected utility of the set of dwelling-type choices available in the neighborhood (H) • alternative specific constants (H)	• non-minority households are less likely to choose locations with concentration of minorities • few of the hypothesized attributes have significant coefficients and intuitive signs
Deng, Ross, and Wachter (2003)	Philadelphia Sample size not specified	Housing tenure and location choices	NL (segmented by race and tenure)	Traffic analysis zones	• size variable (N) • variation in price (N) • racial composition (N) • income composition (N) • zonal fixed effects of housing price (N) • zonal equity risk (N) • geographic indicator (N)	• for white owners occupants, unmarried individuals without a high school education avoid locations with high concentrations of blacks, expensive quality adjusted housing prices, and high equity risk. they also prefer central city over suburban locations • as education level rises, white homeowners loose their aversion to locations with high minority concentrations, high price levels and high equity risk. • black households are more likely to reside in locations with a high percentage of minorities and the effect increases with education • black renters are more likely to live in locations with a high percentage of minorities and less likely to live in locations with equity

						risk • the likelihood of residing in zones with high amenity levels increases with education level • employment access makes a location more attractive, but its affect falls with the probability of unemployment
Earnhart (2002)	Town of Fairfield 105 households residing in own single-family dwellings	Housing choice	MNL / Conjoint	Privately owned residential single family dwellings	• style (H) • #bedrooms (H) • #bathrooms (H) • interior space (H) • lot size (H) • age of structure (H) • geographic indicators (N) • flood frequency (N) • natural features (N) note: extent of n not identified	• households are more likely to choose styles other than cape cod, colonial and ranch-style • households are more likely to buy houses exposed to 100-year flood • households are less likely to locate near the beach • other factors are statistically insignificant note: (1) several counter-intuitive results, perhaps due to lack of appropriate interaction terms; (2) stated-preference data gave more intuitive results
Freedman and Kern (1997)	Philadelphia, Chicago, San Francisco, Detroit and Houston Households headed by non-black, married working couples from the 5% PUMS data (exact sample size not reported)	Residential and work locations	Joint MNL	City vs suburb	• presence of infant (H) • presence of children of any age (H) • household income (H) • husband's and wife's estimated commute time (N) • husband's and wife's projected wages (N)	• commuting times have negative impacts for both spouses in all cities • impact of commute time is always larger for wives than husbands • presence of children reinforces the impacts of commute for wives but not husbands • residential attractiveness of suburbs exceeds that of the city for all two-worker households and the disparity increases consistently with household income, but only sporadically with the presence of children
Gabriel and Rosenthal (1989)	Washington, DC About 2000 households	Residential location	MNL (stratified by race)	Counties	• geographic indicators (N)	• elevating black socioeconomic status to that of white households would alleviate only a small portion of pervasive racial segregation • propensity of a white household to locate in an area increases monotonically with the representation of white households • white location patterns are much more sensitive to changes in socioeconomic status than are those of black households
Horowitz (1995)	Washington, DC area Households that have chosen to own one car and to live in owner-occupied, single-	Residential location and commute mode choices	Joint MNL	Census tracts	• shopping accessibility by modes (N) • racial composition (N) • residential density (N) • per-pupil expenditure for households with children and tracts outside of dc (N) • income dissimilarity (N)	• households prefer less costly housing • segregation by race and income

	family dwellings				• geographic indicator (N) • size variable (N)	
Hunt, McMillan and Abraham (1994)	Calgary 377 stated preference observations	Residential location	MNL	Hypothetical housing alternatives	• money cost per month (H) • number of bedrooms (H) • commute time (N) • travel time to a shopping center (N) • light rail transit (LRT) station (not) within walking distance (N)	• travel time for work trips is more than twice as important as the equivalent time for shopping trips • smaller households tend not to place as high a value on larger dwellings • perceived importance of being within walking distance of LRT by households located within walking distance of LRT in reality is more than twice as high as that by other households
Lerman (1975)	Washington, D.C. 170+ households	Residential location, housing type, car ownership, and commute mode choice	MNL	Census tracts	• housing structure (H) • housing cost (H) • size variable (N) • commute time (N) • income dissimilarity (N) • racial composition (N) • residential density (N) • per pupil school expenditures (N) • municipal and property tax (N) • geographic indicator (N) • generalized shopping price by mode (N)	• households generally prefer areas with low density and high school expenditure • single-worker households are more affected by the level of shopping accessibility than multi-worker households • aversion of whites to non-whites is greater for multi-worker households than single-worker households • income segregation is more pronounced for single-worker households than multi-worker households
Levine (1998)	Minneapolis Region 7547 households containing at least one worker; segmented into 6 groups by income and #workers in household	Residence	NL (community cluster, community) tested against 'flat' logit (community)	Communities defined based on whole or part of Census Places Individual choice set defined by those within an hour's travel radius of the workers' point of work and also within the affordability Communities are grouped into 4 clusters based on residential density and commute time	• commute time for worker 1 (N) • commute time for worker 2 (N) • # housing units (N) • # housing units boarded up (N) • # housing units occupied by owners (N) • median housing price (N) • % school finance generated locally • urban/suburban dummy (N) • density (N)	• except for the low- and moderate-income, single-worker households, the choice processes are better represented by the nested model • housing price does not constitute a barrier to the high-income group's locational decisions • low income households are more sensitive to long-commute • low-to-moderate-income, single-worker households are most attracted to added density plus affordable housing (effects of jobs-housing balance)
Nechyba and Strauss (1998)	Camden County, New Jersey Sample size not specified	Residential location	MNL	School districts	• per pupil school spending (N) • effective property tax rates (N) • crime rate (N) • distance from CBD (N) • %commercial land use (N) • median #rooms (N)	• both local school spending and community entry price are significant determinants • increase in commercial activities and distance from the CBD raise the probability a community is chosen • increase in crime rate decreases the

					• %houses built since 1980 (N) • marginal price of an additional room (N) • private goods consumption (N)	probability a community is chosen • higher housing price decreases the probability a community is chosen • higher housing quality increases the probability a community is chosen
Quigley (1976)	Pittsburgh metropolitan area ~3000 renter hosueholds	Housing choice	MNL segmented a household income and size	Predefined housing types	• housing structure (single detached/apartment/common wall) (H) • #bedrooms (H) • built before 1930 (H) • effective monthly housing cost (H) • housing stock (H)	• lower and middle income rental households are strongly influenced by relative price • larger families are more responsive to relative prices • preference for more bedrooms increases with family size and income level • families with three or more members and families of higher incomes prefer single detached units • structure age does not appear as an significant factor
Quigley (1985)	Pittsburgh metropolitan area 584 recent mover renter households	Choices of dwelling, neighborhood, and town	NL	Census tracts	• structure type (H) • age of structure (H) • condition of structure (H) • #bathrooms (H) • #bedrooms per person (H) • monthly income minus rental payment (H) • proportion of homeowners (N) • median rent (N) • auto commute time (N) (town) • transit commute time (N) (town) • racial composition (N) (town) • size measure (N) • school expenditures per student (town) • municipal expenditures per household (town)	• single detached dwellings are preferred to duplexes; both are preferred to apartment dwellings • households prefer more income and lower housing prices • households prefer more space and additional baths • fraction of homeowners and median rent are not significant determinants • a given reduction in auto commute time has three or four times the effect of an equivalent reduction in compute time by transit; the magnitude of the effect is higher at the neighborhood level than at the town level • compared to white households, black households are more likely to choose neighborhoods of larger fraction of blacks; this racial differentiation is not observed at the town level • households prefer to live in towns where school expenditures and other public expenditures are lowers (outside central city) • the implicit value of accessibility (as a fraction of wage) computed without using the IIA assumption is three times higher than that computed with the assumption
Rapaport(1997)	Tampa, Florida	Residential community, tenure, and housing (price)	Joint discrete-continuous model	Counties	• housing price (N) • government expenditure (N) • population density (N) • school quality (N) note: not all are listed	• incorporating the endogeneity of community choice increases the estimated price elasticity of demand for owner-occupied housing • incorporating the endogeneity of community choice decreases the differences between white and non-white households' demand

Sermons and Koppelman (1998)	Portland 2149 households	Residential location	MNL (with and without factor scores)	Census tracts	• size variables (N) • commute times (N) • geographic indicators (N) • crime (N) • ethnicity composition (N) • socioeconomic status (median home value, median household income, %adult with college degree, % professional workers) (N) • family status (%single family homes, %owner occupied homes, average #rooms, density, average household size, %population over 18-year-old, %married couple households) (N)	• crime is a determinant only for households with children • no gender disparity in sensitivity to commute time • households with children and dual-head households have stronger taste for family-oriented tracts • more affluent households seek out tracts with more affluent populations • racial segregation (white vs. non-white)
Sermons and Koppelman (2001)	San Francisco Bay Area 1307 households with one female and one male worker	Residential location	MNL	Census tracts	• commute time (N) • housing costs (N) • size variables (N) • racial composition variables (N)	• differences between female and male commute time sensitivity is the most profound in households with children • households with a male professional worker, a female non-professional worker, and where the female changed her workplace after the last residence change show the largest gender disparity in commute time sensitivity
Sermons and Seredich (2001)	San Francisco metropolitan	Residential location and auto ownership	Joint MNL	Clusters of TAZ	• expected residential utility of choosing a cluster • alternative specific constants (N)	• the mixed MNL and NL specifications did not reject the MNL model
Tu and Goldfinch (1996)	Lothian region Total of 608 households (single young-person households, young couple households, and households with dependent children)	Sectoral housing submarket and dwelling	Sequential models of MNL structure	Housing sectors	• price/income ratio (H) • dwelling type (H) • average #bedrooms (H) • average age of dwellings (H) • kitchen type (H) • central heating (H) • private garden marketability of sectoral and neighborhood submarkets (N) • commute distance (N) • access to shopping centers (N) • school quality (N)	• single young-person and households with dependent children prefer lower price/income ratio • housing marketability is consistently preferred • households with children prefer large houses • single young-person households prefer to live near shopping centers, while young couples prefer to live far from the shopping areas • large kitchen and central heating are preferred by all types of households • single-person households are indifferent about presence of garden
Wadell (1992)	Dallas and Tarrant counties, Texas 2000 Single full-time worker households from each of the non-	Residence, workplace, and housing tenure	Joint MNL	Census tracts	• commute time (N) • travel time to Dallas cbd (N) • travel time to fort worth CBD (N) • % rental/owner housing supply (N) • mean age of housing units (N) • average #bedrooms (N) • % housing units boarded up (N) • % population by race (N)	• despite higher average levels of education than Hispanics, blacks are much more likely to live in racially isolated neighborhoods, to face more resistance in the workplace and consequently to commute farther to work than other racial groups • distaste for commute is balanced with a preference for larger homes and lower

	Hispanic white, black and Hispanic racial groups				• population density (N) • % (non) family households • employment density (N)	densities farther from the city centers
Waddell (1993)	Dallas and Tarrant counties, Texas households with only one full-time worker who are white non-Hispanic	Residential location and workplace	Joint MNL vs. NL	Census tracts	same as above	• joint specification outperforms the nested structure • higher income and presence of children increase the preference for larger housing • segregation by socioeconomic status, stage of life cycle and ethnicity • housing is treated as a normal good in that more (bigger) is better • high population density is preferred, but not employment density
Waddell (1996)	Honolulu, Hawaii Single and dual-worker households	Housing mobility, tenure, and location	NL	Transport Analysis Zones (TAZ)	• employment accessibility (N) • population accessibility (N) • median housing cost/income ratio (N) • income dissimilarity (N) • size variable (N)	• increase in commute time deduces the probability of a neighborhood being selected • renters are more likely to choose a location that reduces their commute distance • females are more likely to change jobs after a residential move to have a shorter commute • the negative effect of employment accessibility is higher for dual-worker than for single-worker renter households • renter households prefer locations with lower population accessibility (density) • housing cost-income ratio has negative effect on location choice for dual-worker households but not for single-worker households
Weisbrod, Lerman and Ben-Akiva (1980)	Minneapolis/St. Paul 487 households	Residential location, housing type and auto ownership	NL?	Zones	• average housing price by type (H) • stock of various building types and housing unit sizes (N) • income composition (N) • demographic composition (N) • taxes (N) • crime (N) • teacher/pupil ratio (N)	• a 5% reduction in auto commute time has an effect equivalent to 1.5% reduction in monthly rent, 3.8% decrease in home value, or 28% decrease in crime rate per capita • a 5% reduction in commute time by bus has an effect equivalent to 0.3% reduction in monthly rent, 0.5% decrease in home value, or 3.8% decrease in crime rate per capita • teacher/pupil ratio was not a statistically significant determinant • no reduction in commute time could compete with locational effect caused by the propensity of households with children to choose single-family, detached housing

2.6.1 Choice Dimensions and Model Structures

The choice of residential location is very complex and is interdependent on many other choices. For example, a decision to live in an apartment restricts one's choices of location to areas where there are apartment buildings. Hence, the choice of whether to live in a house or an apartment influences the choice of residential location. Similarly, a worker who prefers to commute by transit would choose to live near a transit stop. Causality can also operate in the other direction. A worker who moved to a neighborhood for non-transport related reasons may subsequently decide to commute by transit just because the mode is available in the new neighborhood. The recognition of this interdependency among households' long-term choices has lead researchers to model residential location choice jointly with other choice dimensions, such as work location (Abraham and Hunt, 1997; Freedman and Kern, 1997; Wadell, 1992, 1993), commute mode (Abraham and Hunt, 1997; Anas, 1981; Anas and Chu, 1984; Horowitz, 1995; Lerman, 1975), car ownership (Lerman, 1975; Sermons and Seredich, 2001; Weisbrod, Lerman and Ben-Akiva, 1980), housing mobility (Clark and Onaka, 1985; Waddell, 1996), housing tenure (Boehm, 1982; Deng, Ross, and Wachter, 2003; Rapaport, 1997; Wadell, 1992, 1996), specific housing attributes (Boehm, 1982; Clark and Onaka, 1985; Lerman, 1975; Rapaport, 1997; Tu and Goldfinch, 1996; Weisbrod, Lerman and Ben-Akiva, 1980), and travel propensities (Ben-Akiva and Bowman, 1998). The exact combination of choice dimensions modeled in each study is listed in column 3 of Table 2.1.

As identified in column 4 of Table 2.1, it is common for the aforementioned studies to adopt the MNL structure and to incorporate multiple choice dimensions by forming a large, but still finite, super choice set as the Cartesian product of one choice set with the other. For example, if one were modeling commute mode choice and residential location choice jointly, all the possible combinations of mode and residential location would become the full choice set. The

key disadvantage to this approach lies in the fact that it can vastly enlarge the choice set and care must be exercised to eliminate infeasible combinations from this large choice set (Lerman, 1983). A popular alternative to treating the residential location choice along with other choice dimensions is to apply the NL model such that one of the levels in the nesting structure corresponds to the residential location choice. Irrespective of the model structure used (joint MNL or NL), all of the joint choice studies cited in Table 2.1 address the residential location aspect of the choice behavior based on Lerman's grouped alternative approach.

Given the complexity behind households' decision making about residential location and other choices, it is certainly not a simple task to accommodate the complete interplay among all the relevant choice dimensions in a single model. This is evident in the fact that most of the aforementioned joint choice modeling efforts consider at most two or three choice dimensions and treat other choice dimensions as given. With the focus of building a more sophisticated specification for the residential utility than that found in the joint-choice models, other studies focus solely on the residential location choice. These studies include Chattopadhyay (2000), Earnhart (2002), Gabriel and Rosenthal (1989), Hunt, McMillan and Abraham (1994), Levine (1998), Nechyba and Strauss (1998), Quigley (1976), Quigley (1985), Sermons and Koppelman (1998), and Sermons and Koppelman (2001). With the exception of Hunt, McMillan and Abraham (1994) and Earnhart (2002) who considered each dwelling as a choice alternative, these studies all assumed the grouped alternatives model and consider groupings of dwellings, defined by either physical proximity or other housing attributes, as choice alternatives.

2.6.2 Definition of Residential Choice

Evidently, despite the IIA property inherit from its MNL structure and the estimation bias introduced by using average values of dwelling attributes to describe the groupings, Lerman's grouped alternatives approach has been applied extensively in previous studies of residential

location choice. This use of the grouped alternatives model to approximate the ideal disaggregate models was once forced on the analyst simply because data are not available for all the alternatives at the original level of the elemental alternatives (Lerman, 1983). Yet, although over the years more micro-level data has become available, residential location choice studies of a disaggregate nature remain scarce. This is perhaps because few researchers have risen to challenge the norm, *i.e.* the aggregate approach, or because the concept of grouped alternatives has its behavioral merits.

As pointed out in Section 1.2, defining choice alternatives and choice set for studies of spatial choices is often not as straightforward as for those of non-spatial choices. In the context of residential location choice in a metropolitan area, the number of potential alternatives is typically large. However, in reality it is highly unlikely that households consider each and every available housing unit. Rather, they devise their own search strategies and form their own choice sets based on the resources available to them and their limited capacity for gathering and processing information. Ideally, for modeling purposes, we would like to identify the choice alternatives and the choice set as perceived by individual households. Yet the task is a difficult, if not an impossible, challenge to the analyst (Kanaroglou and Ferguson, 1998). Instead, analysts of the housing market have treated the problem of potentially large choice set by representing the heterogeneity of the dwellings and neighborhoods available to consumers by a small number of groupings of residential housing. The groupings, however, are often arbitrarily defined.

As shown in column 5 of Table 2.1, the most common practice is to aggregate alternative dwellings into administratively defined units, typically census tracts or transport analysis zones. The tracts or zones are then considered as the *communities* or *neighborhoods* that the individual households choose from. Other administratively defined units used as proxy for residential alternatives include counties (Gabriel and Rosenthal, 1989; Rapaport, 1997), school districts

(Nechyba and Strauss, 1998) and census cities (Levine, 1998). This use of administrative units is likely attributed to the fact that spatial data describing the residential environment of the dwellings are often readily available only for these units. The only studies found to use non-administratively defined units are Anas (1981) and Anas and Chu (1984), in which the analysts divide the study area into 0.5 by 0.5 mile-square zones. Data about the housing quality and neighborhood amenity are then aggregate and/or disaggregate over these "quarter-sections".

Instead of using locality-based groupings, some studies construct residential choice alternatives by grouping individual housing units based on their non-spatial attributes. For example, Quigley (1976) and Tu and Goldfinch (1996) both defined different housing types as choice alternatives. In examining individuals' preferences for neighborhood qualities, Boehm (1982) and Chattopadhyay (2000) considered as choice alternatives the neighborhood types defined based on tract-level median income values, as opposed to the individual neighborhoods. Similarly, in Levine (1998), communities defined based on census places are grouped into 4 clusters based on residential density and commute time to form the location choices for individuals.

2.6.3 Measurement of Spatial Factors

Undoubtedly, for any residential choice analysis to accurately reflect the underlying choice utility, it is important to include all the key choice variables in the model specification. Equally important to the analysis is to introduce these variables at the appropriate scale, especially for the spatial variables.

In previous studies that employ the grouped alternatives model, the scale at which the spatial variables are introduced typically follows from the spatial definition of the choice alternatives. That is, as listed in column 6 of Table 2.1, the neighborhood variables considered in these studies are measured over the same spatial unit as those defined as the residential choice

alternatives (identified in column 5 of the same table). For instance, if census tracts are used as the residential alternatives, averages of the unobserved individual dwelling attributes (denoted as $\overline{X}_{n,c}$ in Equation (2.15)) and measures of the public amenity and other characteristics of the surrounding neighborhood (denoted as $Y_{n,c}$ in Equation (2.15)) are constructed accordingly for the census tracts. These studies all assume that such constructs provide an accurate representation of the residential neighborhoods or communities as perceived by the decision makers. The effects of neighborhood characteristics (for example, affluence) are inferred directly from the estimates obtained for the corresponding parameters measured over the census tracts or TAZ (for example, zonal medium income).

2.6.4 Significant Choice Determinants

A wide variety of dwelling unit attributes and neighborhood attributes have been empirically shown to influence residential choice behavior. Some of these attributes are found to have different effect on households of different characteristics. Some attributes are found to have significant influences in certain studies, but insignificant in others. Findings from previous residential choice studies about the various choice variables (listed in column 7 of Table 2.1) are summarized below.

Commuting

A hypothesis often tested, and proven to be true over and over again, is that, all else being equal, households prefer residence resulting in lower commuting time/cost. The disutility associated with commuting has been found as one of the dominant influences on residential choice in previous studies for different cities and population segments. Moreover, several studies have shown that households of different characteristics exhibit varying degrees of sensitivity to commuting.

Studies that found greater sensitivity to commute time for females relative to males (irrespective of being married or single) include Waddell (1996), Abraham and Hunt (1997), and Sermons and Koppelman (2001)[1]. This gender disparity is often attributed to females' household responsibilities: "[w]omen who work must also keep house, cook dinner, and perhaps be at home when their children arrive from school. Thus women are thought to place a higher value than men on time spent commuting, and they choose either their jobs or their residences so as to shorten the journey" (White, 1977). This is supported by the finding in Freedman and Kern (1997) that the presence of children reinforces the impacts of commute for wives but not husbands. Another explanation for the gender difference is that working women often are secondary wage earners who take a more casual attitude toward job seeking than that of the primary wage earner (Kain, 1962). They therefore choose workplaces that are close to their residence. An empirical evidence is offered by Sermons and Koppelman (2001), who found that households with a male professional worker, a female non-professional worker, and where the female changed her workplace after the last residence change show the largest gender disparity in commute time sensitivity.

Commute time has also been found to have more effect on low- income than high-income households (Levine, 1998). The finding is said to be in contrast to a higher value of commute time for high-income households as found in modal choice models. "[I]n short-run, modal choice decisions, individuals' income may influence their time/money tradeoffs, but when it comes to longer-term choices such as where to live, all people find themselves within the same twenty-four-hour day. Cycles of work, leisure, and sleep may then become more important than opportunities to save some time by spending some money, or vice versa" (Levine 1998).

[1] Gender disparity in sensitivity to commute time is found to be insignificant in Sermons & Koppelman (1998).

Access and Accessibility

In addition to commuting, which indicates the ease of access to employment, previous studies provide results about how access to other type of opportunities, or land-use, affects residential choice behavior.

After accounting for the effect of commuting, Waddell (1993) found that households do not prefer high employment accessibility. Furthermore, the negative effect of employment accessibility is higher for dual-worker than for single-worker renter households. Nechyba and Strauss (1998) showed that, as the amount of commercial activities increases in a zone, the probability of that zone being chosen increases. The propensity for easy access to shopping opportunities is found to be only half as important as access to workplaces (Hunt, McMillan and Abraham, 1994). Tu and Goldfinch (1996) found differing effect of access to shops. Single young-person households prefer to live near shopping centers, while young couples prefer to live far from the shopping areas.

Access to alternative modes of transportation can also have an effect on residential choice. As Hunt, McMillan and Abraham (1994) found in their stated preference analysis, residents who already live within walking distance of light rail transit perceive the ease of access to transit more than twice as important as other households do.

Residential density

Conflicting observations have been drawn about the effect of residential density on residential choice behavior. While Lerman (1975), Horowitz (1995) and Ben-Akiva and Bowman (1998) showed that households generally have an aversion to locations with high density, Waddell (1993) and Rapaport (1997) found that high population density is preferred by households. The contradictory findings may be attributed to the difference in the population

segment and/or the geographical area being studied. It could also be a result of other sources of error such as omitted variables or aggregation bias.

Disparity in the effect of density has also been found for different population groups. Lower density is especially attractive for large households, while low-to-moderate-income households and single-worker households are most attracted to higher density (Levine, 1998). Moreover, all else being equal, home-owners (according to Waddell, 1996) and African-American households (according to Waddell, 1993) are more likely to locate in areas of high density.

Housing price and quality

Housing affordability, measured by housing price (as in Nechyba and Strauss, 1998) or by price-to-income ratio (Quigley, 1985; Waddell, 1992; Levine, 1998; and Ben-Akiva and Bowman, 1998), is generally found to be an attractive feature for a residential zone. However, housing price does not seem to constitute a barrier to the residential decisions of high-income households (Quigley, 1976; Levine, 1998) and single-worker households (Waddell, 1996). As suggested by Quigley (1976), larger families are strongly influenced by relative price. Tu and Goldfinch (1996) found similar results for single young-person households and households with dependent children. Waddell (1992) found a positive, though small, elasticity for housing price for white workers. This is contradictory to the observation made by Deng, Ross and Wachter (2003) that white homeowners loose their aversion to locations with high price levels as education level rises.

The propensity of dwelling size varies by household size, family status, socioeconomic status, and gender. Quigley (1976), Hunt, McMillan and Abraham (1994), and Chattopadhyay (2000) all found that a larger family size increases the probability of choosing a large unit. Quigley (1976) and Waddell (1993) found preference for more bedrooms increasing with income

level. Presence of children also increases the preference for larger housing (Waddell, 1993; Tu and Goldfinch, 1996). Somewhat counterintuitive results were found in Chattopadhyay's (2000) study, which showed that Caucasian households opt for less number of rooms, older houses, and bigger lot size than non-whites.

The preference for housing structure type has been examined in a couple of studies. Preference for single detached housing is found for families with three or more members and families of higher incomes (Quigley, 1976). The propensity for single-family, detached housing is also found of households with children (Weisbrod, Lerman & Ben-Akiva, 1980). Furthermore, Quigley (1976) found that, while single detached dwellings are preferred to duplexes, both are preferred to apartment dwellings.

Other housing attributes identified to have positive effect on residential choice behavior include the presence of large kitchen, central heating and garden (Tu and Goldfinch, 1996), and architecture styles other than cape cod, colonial and ranch-style (Earnhart, 2002). Age of housing structures, on the other hand, does not appear to be a significant factor (Quigley, 1976).

Race and ethnicity

As indicated by the parameter estimates associated with racial composition variables interacted with the race of the sampled households, households often show strong preference for locations with a high percentage of households of the same race (for example, Lerman, 1975; Clark and Onaka, 1985; Quigley, 1985; Horowitz, 1995; Sermons and Koppelman, 1998; Chattopadhyay, 2000). Past theoretical and empirical studies have suggested a number of theories to explain racial segregation. One hypothesis is that segregation results from economic differentials among racial groups. However, as shown in Gabriel & Rosenthal (1989), at least for the Washington DC area, elevating black socioeconomic status to that of white households would alleviate only a small portion of pervasive racial segregation.

Other theories of segregation include racial discrimination in lending and housing markets and differences between racial groups in preferences for neighborhood attributes. The effect of these two factors may have been absorbed into the observed effect of other attributes. For instance, Deng, Ross, and Wachter (2003) found that, as education level rises, white homeowners loose their aversion to locations with high minority concentrations. Black households, on the other hand, are more likely to reside in locations with a high percentage of minorities and the effect increases with education. In Wadell (1993), despite higher average levels of education than Hispanics, blacks are found to be much more likely to live in racially isolated neighborhoods than Hispanics.

Interestingly, Quigley (1985) found evidence of racial segregation at the neighborhood (census tract) level, but not at the town level. This finding suggests different choice of spatial units used for measuring a choice determinant may lead to very different results and interpretations about the effect of that determinant on the residential choice behavior.

Socioeconomic status

Segregation by socioeconomic status, which may have in part contributed to the racial segregation observed in some studies, has been found by, among others, Horowitz (1995) and Sermons and Koppelman (1998). Other studies also found interdependency between socioeconomic status and race. Both Boehm (1982) and Chattopadhyay (2000) indicated that blacks are less likely to live in high-income neighborhoods. Also, white households are much more sensitive to changes in socioeconomic status than are black households (Gabriel and Rosenthal, 1989).

Age and family status

In addition to segregation by race and socioeconomic status, households also tend to cluster in areas with other households of similar age and family structure. Among others,

Waddell (1992) and Sermons & Koppelman (1998) have shown that households choose locations with higher proportions of the same family structure.

School quality

It is intuitive that school quality should play an important role in the residential location decision, especially for families with children. Yet, empirical findings about the effect of school quality have been mixed. Lerman (1975), Nechyba and Strauss (1998) and Chattopadhyay (2000) found positive effects of school quality on residential choice utility. However, Ben-Akiva and Bowman (1998) found school quality to be an insignificant factor and Quigley (1985) indicated that school expenditures have a negative influence on residential choice utility. These inconclusive findings may be attributed to inaccurate measurement of school quality and presence of correlation between school quality and other factors.

Safety

Where considered, the safety of residential neighborhoods is often captured in residential choice models by the observed crime rates. Both Ben-Akiva and Bowman (1998) and Nechyba and Strauss (1998) found that households are less likely to locate in areas with high crime rate. Measured at the town level in the study by Sermons & Koppelman (1998), crime rate is found to have negative influence only for households with children.

Alternative specific constants

In addition to the aforementioned choice determinants whose effects have been empirically shown, there are many other factors which are subjective and personal (such as the view from a property or the cleanliness of a street) and are obviously difficult to introduce in any quantitative analysis. In theory, the average effects of the factors unobserved by the analyst can be captured by introducing alternative specific constants (ASC) into all but one utility function,

which would act as a base against which the effects of the other variables are measured. In practice, however, since residential alternatives are unranked and often numerous even after the grouping process, past studies usually do not use ASC. The exception is when there are relatively few alternatives (for example, the choice is either urban or suburban) or when all or some of the location alternatives have well-known geographical identities (for example, when the choice set consists of counties, or when some of the choices correspond to the CBD). In the latter case, dummy indicators can be used to capture the unobserved effects specific to these geographical areas. For instance, Chattopadhyay (2000) used the CBD indicator to isolate the unobserved negative effect of CBD living on large and white families.

2.7 Summary

This chapter has described the conventional approach to modeling residential location choice using the discrete choice framework. Decision makers' choice behaviors are assumed to follow the RUM principle and are subsequently modeled using the MNL or NL formulations. Because the number of choice alternatives is often large in residential choice problems and data about the elemental alternatives are often unavailable, the application of the MNL and NL models are typically coupled with the use of grouped alternatives. The popularity of this conventional approach is evident in the literature survey presented in this chapter of past modeling efforts of residential location choice. The survey identified the differences and commonalities shared among these studies. As will be discussed in the next chapter, most of these studies are flawed in that they do not appropriately address the spatial nature of the choice problem.

CHAPTER 3

SPATIAL COMPLEXITIES IN MODELING RESIDENTIAL LOCATION CHOICE

As revealed in the preceding chapter, the logit family of models has been the standard tool for studying individual's residential preference since its conception in the 70's. There has not been any major advance in the conceptualization of, or the methodology for treating, the spatial choice problem. Of the four spatial issues identified in Section 1.2, the analysts typically circumvent the first two issues (namely, definition of choice alternatives and definition of choice set) by treating the choice problem as one of selecting from groupings of elemental alternatives (*i.e.* the individual dwellings). Characteristics of the elemental alternatives are aggregated, or averaged, to represent the characteristics of the grouped alternatives. The logit models are then applied in the same manner as for non-spatial contexts, with the remaining spatial issues assumed away.

The aim of this chapter is (1) to establish the significance of the two spatial issues; substitutability among choice alternatives and measurement of spatial factors; that are typically unaddressed in residential location choice models; and (2) to explore possible ways of treating these issues. The chapter is divided into two sections. Section 3.1 focuses on the substitutability problem and explains why the residential location choice problem is very susceptible to such a problem. A number of modeling techniques that may be used to resolve the problem are also discussed. Section 3.2 explains the issues, and their origins, related to the representation of spatial factors. It also discusses possible ways of operationalizing the concept of neighborhood to improve the representation of spatial factors.

3.1 Substitutability among Alternative Locations

Following directly from the assumption that the error terms are independently (no correlation) and identically (same variance) distributed, the IIA property that characterizes the MNL model structure implies proportional substitutability among choice alternatives, which is a very restrictive condition. In cases where the condition is violated, the mis-representation of choice behavior will results in biased estimates and incorrect predictions of likely future behavior. The subsequent section explains why, from a behavioral standpoint, the residential location choice problem is very susceptible to violations of IIA. The explanation is followed by Section 3.1.2, which briefly reviews a number of modeling techniques that may be used to accommodate more flexible substitution patterns.

3.1.1 Theories of Choice Substitutability

In situations such as residential location choice, the number of potential alternatives could be huge. Because of limits on our ability to process information, it is not likely that individuals evaluate all the alternatives as when the choice set is small (Thill, 1992). Instead, they process spatial information and make spatial choices in such a way that they first evaluate clusters of alternatives and only evaluate particular alternatives from within a selected cluster (Fotheringham et al, 2000). For example, an individual searching for housing in a city might have strong feelings about which parts of the city she or he would like to live in and which parts are to be avoided. The feelings might be attributed to personal residential experience in the past or the sense of attachment to family and friends. This behavior of hierarchical information processing has been hypothesized and supported by, among others, Hirtle and Jonides (1985), McNamara (1992) and Fotheringham and Curtis (1999).

The hierarchical choice behavior may be modeled by the NL formulation if the clusters of alternatives are well defined. For example, if it is well established that households indeed select

first a community (or neighborhood) and then a dwelling unit, and that the communities have unambiguous geographic definition, then the NL models can be applied with relative confidence. Alternatives in a community face a more competitive choice context than do alternatives not in the same community. The NL formulation, in this case, provides a better approximation for the hierarchical information process than the MNL model. However, the reality is that boundaries of people's mental clusters might be physical, perceptual or entirely subjective (Fotheringham and Curtis, 1999). Individuals themselves might not be able to tell how they process spatial information and the mental clusters may be fuzzy rather than discrete (Fotheringham et al, 2000). It is therefore very difficult, if not impossible, for the analyst to construct nests of alternatives that accurately reflect these mental clusters. The imposition of artificial boundaries introduces problematic situations such as two nearby alternatives that are separated into different clusters being considered as less substitutable than two alternatives located at two far ends of the same cluster.

Even if individuals' mental clusters could be identified, one still cannot not assume that alternatives within the same cluster are equal substitutes of one another (Fotheringham, 1988). This is because alternative choice locations are spatially dependent. That is, they have fixed, unique geographical relationships with one another. The locations that are close to each other are more likely to share similarities than those that are farther apart. It is up to the analyst to include in the utility specification all the variables that capture these similarities. However, certain housing and environmental attributes are by nature subjective and difficult to measure (for example, sentimental attachment to the area, the cleanliness of a street, the view from a property or other aesthetic characteristics). The similarities attributed to such attributes are usually not explicitly incorporated into the model and would therefore result in correlations among the

stochastic utilities. This spatial correlation results in the violation of IIA and, consequently, the unproportional substitutability among choice alternatives.

3.1.2 Models with Flexible Substitutability

When IIA is not satisfied, Train (2002) suggests the analyst to take one of the following approaches: (1) use the logit model under the current specification of deterministic utility while considering the model to be an approximation; (2) re-specify the deterministic utility so that the source of the correlation is captured explicitly and thus the remaining errors are independent; or (3) use a different model structure that allows for correlated errors. The first approach obviously does not serve the purpose of this dissertation research. The second approach requires further investigation into, for example, how individuals process and perceive the various choice characteristics. This approach will be examined in Section 3.2. For now, let us assume that attributes about the elemental alternatives are unavailable, as in most previous studies, and so one cannot avoid using aggregate zones as choice alternatives. In this case, correlation is inevitable and the third approach of using an alternative model structure must be taken.

Since the MNL model was first developed, there have been significant efforts devoted to develop models that allow more flexible substitution patterns for the alternatives. These models are described below.

3.1.2.1 The probit model

The probit models are derived from a relaxation of the IID assumption such that the unobserved error components are assumed to be jointly normally distributed with a general variance-covariance matrix. That is, the error component can have a different variance for each alternative and can be correlated across alternatives. The probit choice model is given by substituting the normal density function into Equation (2.2):

$$P_{n,i} = \int_{\varepsilon_n} I\left(\varepsilon_{n,j} - \varepsilon_{n,i} < V_{n,i} + V_{n,j}, \forall j \neq i\right)\Phi(\varepsilon_n)d\varepsilon_n . \quad (3.1)$$

The probit model structure is powerful in its capability to accommodate flexible substitution patters. However, the power comes at the expense of certain theoretical and computational issues (see Daganzo, 1979, for a detailed discussion). One issue relates to the large set of model parameters needed to estimate due to the unknown variance and covariance terms. The covariance parameters generate conceptual problems relating to the difficulty in interpreting the covariance parameters and in forecasting the effects of introducing new alternatives (Horowitz, 1991). Moreover, the evaluation of the multivariate normal integrals on the right hand side of Equation (3.1) is computationally difficult. Although development of simulation based approaches (for example, Börsch-Supan and Hajivassiliou, 1993) has alleviated the computational difficulty to some degree, application of probit models is still limited. Bolduc (1992), Bolduc, Fortin and Gordon (1997) are among the few applications of multinomial probit models in a spatial context. In order to reduce the number of nuisance parameters to estimate in the error covariance matrix, both studies introduced an autoregressive process based on a distance decaying relationship. Since the number of alternatives considered in their empirical analysis was only 18, it is unclear if it is computationally feasible to estimate the model for much larger choice sets, as typically found in the residential location context.

3.1.2.2 The mixed multinomial logit model

The mixed multinomial logit (MMNL) models, also known as random-coefficients logit models, are a generalization of the logit models (Bhat, 2003). They are motivated by the desire to allow the model parameters to vary across the population rather than being fixed. That is, each decision maker assigns different coefficients, β_n, to the observed variables, $X_{n,j}$:

$$P_{n,i} = \frac{e^{\beta_n' X_{n,i}}}{\sum_j e^{\beta_n' X_{n,j}}} \tag{3.8}$$

Unless replications of each decision maker's choices are observed, β_n is not estimable. Instead, the coefficients β_n are assumed to follow a random distribution with a density function $g(\beta \mid \theta)$, where θ are underlying moment parameters of $g(\cdot)$, and integrate the logit probability over β_n to obtain:

$$P_{n,i}(\theta) = \int_{-\infty}^{+\infty} \frac{e^{\beta_n' X_{n,i}}}{\sum_j e^{\beta_n' X_{n,j}}} g(\beta_n \mid \theta) d(\beta_n) \tag{3.9}$$

One can then proceed to estimate θ, the population parameters describing the distribution of β_n.

In addition to being used to accommodate unobserved taste variations across decision makers, the MMNL model structure can also be employed to capture flexible substitution patterns across alternatives. This is because the decision maker's tastes introduce correlations among the unobserved component of utility over alternatives. The MMNL models do not exhibit the IIA property and can, in fact, approximate any substitution patterns (McFadden and Train, 2000). The MMNL model structure is also conceptually appealing and easy to understand since it is the familiar MNL model mixed with the multivariate distribution (generally multivariate normal) of the random parameters (see Hensher and Greene, 2002). In the context of relaxing the IID error structure of the MNL, the MMNL model represents a computationally efficient structure when the number of error components needed to generate the desired error covariance structure across alternatives is much smaller than the number of alternatives (see Bhat, 2002a).

3.1.2.3 The generalized extreme value model

The generalized extreme value (GEV) class of models, as the name implies, are based on a generalization of the extreme value distribution. The model takes the form of:

$$P_{n,i} = \frac{e^{V_{n,j}} \dfrac{\partial \mathrm{G}\left(e^{V_{n,1}},\ldots,e^{V_{n,J}}\right)}{\partial e^{V_{n,i}}}}{\mathrm{G}\left(e^{V_{n,1}},\ldots,e^{V_{n,J}}\right)} \tag{3.2}$$

where the function $\mathrm{G}\left(e^{V_{n,1}},\ldots,e^{V_{n,J}}\right)$ is non-negative, linear homogenous, tending toward $+\infty$ when any argument tends toward $+\infty$ and has mth cross-partial derivatives which are non-negative for odd m and non-positive for even m (McFadden, 1978). If $\mathrm{G}(\cdot)$ satisfies these conditions, then the function:

$$\mathrm{F}(\varepsilon_1,\ldots,\varepsilon_J) = e^{-\mathrm{G}\left(e^{-\varepsilon_1},\ldots,e^{-\varepsilon_J}\right)} \tag{3.3}$$

is a multivariate extreme value distribution and the above model is consistent with utility maximization. Thus, the GEV model relaxes the IID assumption of the MNL by allowing the random components of alternatives to be correlated, while maintaining the assumption that they are identically distributed (*i.e.*, identical, non-independent, random components).

Different specifications of $\mathrm{G}(\cdot)$ lead to different GEV models. For instance, if one lets:

$$\mathrm{G}\left(e^{V_{n,1}},\ldots,e^{V_{n,J}}\right) = \sum_{j=1}^{J} e^{V_{n,j}}, \tag{3.4}$$

then Equation (3.2) reduces to the logit choice probability as given in Equation (2.4). Alternatively, if the J alternatives are partitioned into K groups labeled as $B_1,\ldots,B_K$ and let:

$$\mathrm{G}\left(e^{V_{n,1}},\ldots,e^{V_{n,J}}\right) = \sum_{k=1}^{K}\left(\sum_{j\in B_k} e^{V_{n,j}/(1-\sigma_k)}\right)^{1-\sigma_k}, \tag{3.5}$$

where $0 \le \sigma_k < 1, \forall k$, then Equation (3.2) becomes:

$$P_{n,i} = \frac{e^{V_{n,i}/(1-\sigma_k)}\left(\sum_{j\in B_k} e^{V_{n,j}/(1-\sigma_k)}\right)^{-\sigma_k}}{\sum_{k=1}^{K}\left(\sum_{j\in B_k} e^{V_{n,j}/(1-\sigma_k)}\right)^{1-\sigma_k}}, \tag{3.6}$$

which is the NL formulation as define by Equation (2.9) (Williams, 1977; McFadden, 1978; Daly and Zachary, 1978).

In addition to the NL model, several other GEV models have been developed over the years. Intended for use with ordinal discrete choices, the ordered GEV model (Small, 1987) allocates each alternative to nests based on their proximity in an ordered set. The paired combinatorial logit (PCL) model (Chu, 1989; Koppelman and Wen, 2000) allows elemental alternatives to belong "fractionally" to multiple nests, as opposed to the unique nest required in the NL model. Each alternative is allocated in equal proportions to a nest with each other alternative to allow different covariances for each pair of alternatives. The cross-nested logit (CNL) model (Vovsha, 1997) allows differential similarities between pairs of alternatives by allocating a fraction of each alternative to a set of nests with equal logsum parameters across nests.

More recently, Wen and Koppelman (2001) proposed a general GEV model structure, which they referred to as the Generalized Nested Logit (GNL) model. The model is derived from the function:

$$\mathrm{G}\left(e^{V_{n,1}},\ldots,e^{V_{n,J}}\right)=\sum_{k=1}^{K}\left(\sum_{j\in N_k}\left(\alpha_{jk}e^{V_{n,j}}\right)^{1/\mu_k}\right)^{\mu_k}, \qquad (3.7)$$

where N_k is the set of all alternatives included in nest k; α_{jk} is the allocation parameter that characterizes the portion of alternative j assigned to nest k (with the condition that $\alpha_{jk}\geq 0$ and $\sum_k \alpha_{jk}=1, \forall k$); and μ_k is the logsum or dissimilarity parameter for nest k ($0\leq\mu_k\leq 1$). The resulting probability function, derived from substituting Equation (3.7) into Equation (3.2), is

$$P_{n,i} = \frac{\sum_{k=1}^{K}\left((\alpha_{ik} e^{V_{n,i}})^{1/\mu_k} \left(\sum_{j \in N_k} \left(\alpha_{jk} e^{V_{n,j}} \right)^{1/\mu_k} \right)^{\mu_k - 1} \right)}{\sum_{k=1}^{K}\left(\sum_{j \in N_k} \left(\alpha_{jk} e^{V_{n,j}} \right)^{1/\mu_k} \right)^{\mu_k}}$$

$$= \sum_{k=1}^{K}\left(\frac{(\alpha_{ik} e^{V_{n,i}})^{1/\mu_k}}{\sum_{j \in N_k} \left(\alpha_{jk} e^{V_{n,j}} \right)^{1/\mu_k}} \frac{\left(\sum_{j \in N_k} \left(\alpha_{jk} e^{V_{n,j}} \right)^{1/\mu_k} \right)^{\mu_k}}{\sum_{k=1}^{K}\left(\sum_{j \in N_k} \left(\alpha_{jk} e^{V_{n,j}} \right)^{1/\mu_k} \right)^{\mu_k}} \right). \quad (3.8)$$

This probability function can be rewritten as

$$P_{n,i} = \sum_{k=1}^{K} P_{i|k} P_k \ , \quad (3.9)$$

where P_k, the probability of nest k, is

$$P_k = \frac{\left(\sum_{j \in N_k} \left(\alpha_{jk} e^{V_{n,j}} \right)^{1/\mu_k} \right)^{\mu_k}}{\sum_{k=1}^{K}\left(\sum_{j \in N_k} \left(\alpha_{jk} e^{V_{n,j}} \right)^{1/\mu_k} \right)^{\mu_k}}, \quad (3.10)$$

and $P_{i|k}$, the probability of alternative i if nest k is selected, is

$$P_{i|k} = \frac{\left(\alpha_{ik} e^{V_{n,i}} \right)^{1/\mu_k}}{\sum_{j \in N_k} \left(\alpha_{jk} e^{V_{n,j}} \right)^{1/\mu_k}} . \quad (3.11)$$

Wen and Koppelman (2001) showed that GNL model is consistent with random utility maximization if the condition $0 \le \mu_k \le 1$ holds for all k. Under the usually linear-in-parameter assumption about the utility function, the direct elasticity of an alterative j is

$$\sum_{k=1}^{K}\left(P_{j|k} P_k \left(\left(1 - P_j\right) + \left(\frac{1}{\mu_k} - 1 \right)\left(1 - P_{j|k}\right) \right) \right) \frac{\beta_q x_{n,jq}}{P_j} \quad (3.12)$$

If the alternative does not share a nest with any other alternative or is assigned only to nests for which the logsum value equals one, the above expression reduces to the MNL direct elasticity stated in Equation (2.6). Also, the cross elasticity of a pair of alternatives, i and j, is

$$-\left(P_i + \frac{\sum_{k=1}^{K}\left(\frac{1}{\mu_k}-1\right)P_k P_{i|k} P_{j|k}}{P_j} \right)\beta_q X_{n,jq}\,. \qquad (3.13)$$

In this case, if the alternatives do not share any common nest, the above expression reduces to the MNL cross elasticity given in Equation (2.7). On the other hand, if the pair of alternatives is in one or more common nests with logsum less than one, the cross elasticity for the GNL model is greater in magnitude than for the MNL model. The model accommodates differential cross elasticity among pairs of alternatives through the fractional allocation of each alternative to a set of nests, each of which has a distinct logsum parameter. The elasticity increases in magnitude as the logsum parameter decreases from one, with the magnitude of the impact related to the probability of the nest and the conditional probabilities of the alternatives in the nest.

Swait (2001) independently proposed the choice set Generation Logit (GenL) model which has a structure similar to the GNL model. In the GenL model, each nest represents a possible choice set so that the marginal probability represents the selection of the choice set and the conditional probability represents the choice of an alternative given that the choice set. The difference between the GenL and the GNL models is that, in the former, the allocation parameters are associated with individuals rather than alternatives.

It has been shown that the GNL includes the two-level NL, the PCL, and the CNL as special cases (Wen and Koppelman, 2001). The GNL formulation is conceptually appealing from a formulation standpoint and allows substantial flexibility. However, in practice, the flexibility of the GNL model can be realized only if one is able and willing to estimate a large number of

dissimilarity and allocation parameters. The net result is that the analyst will have to impose informed restrictions on the model formulation that are customized to the application context under investigation.

3.2 Representation of Spatial Factors

In the residential location choice problem, because of the alternatives being inherently spatial, the variables characterizing the alternatives are, as expected, also spatial in nature. This means that values of these variables can be observed only after a geographic definition has been specified. For variables describing the dwelling and the land lot, the geographic definition is unambiguous and follows from the dwelling structure. For other variables that describe aspects of the surrounding of a given dwelling, the geographic definition is tied to the concept of neighborhood, or community, whose boundaries are usually not clear-cut. This definitional problem has been constantly overlooked in previous studies of residential choice. Almost without exception, neighborhood attributes are measured using administratively boundaries. The approach implicitly assumes that households select administrative areas to live as part of their residential location choice. Such an assumption may be valid if the administration areas are, for example, school districts or counties. However, it would be very unrealistic when census tracts or TAZ are used.

Problem also arises when the grouped alternatives approach is taken and administrative units are used as choice alternatives. The use of the aggregate measures to represent the characteristics of the constituting elemental alternatives introduces aggregation errors into the observed component of the choice utility. The imposition of artificially-defined administrative boundaries also introduces the conceptual problem that alternatives within the same administrative unit are equally substitutable after observable characteristics are accounted for.

In Section 3.2.1, the implication of aggregate measures is discussed in more details, along with a brief review of techniques that could possibly address the aggregation problem. Attention is turned over to the more fundamental problem of how neighborhoods should be spatially defined for analytical purposes in Section 3.2.2, which presents ideas and findings drawn from past studies from various disciplines.

3.2.1 Implication of Aggregate Measures

In quantitative studies, data analysis involves a variety of judgments and decisions. A fundamental consideration is the level of detail at which the analysis takes place. While a variable may either be continuous (such as length) or discrete (such as make and model of a car) in nature, in theory, values observed for the variable can be recorded using any nominated scale. Often, if the variable is continuous or if the values are diverse, the underlying value range is segmented or classified to yield coarser measurement units. Data are then aggregated over these units to reduce the level of detail.

Data that describe attributes of geographical entities, including dwellings, land lots and neighborhoods, is of particular interest to residential choice studies. Previous studies typically involve aggregation of spatially scattered dwellings into predefined zones and use the zones as choice alternatives. During the aggregation process, information is lost about the observed uniqueness of, and the variations between, the dwellings that fall within the same zone. As a study region can be segmented in different ways (in terms of shape, size, and orientation) to yield different zoning systems, the magnitude of information loss may vary. Consequently, the result of further analysis of the data will vary.

The uncertainty as to what spatial units to use has been known to spatial analysts as the modifiable areal unit problem (MAUP) (Openshaw, 1984) that is endemic to all aggregate data. The effect of the MAUP has been found in a variety of spatial analysis and modelling studies,

including univariate statistical analyses (Gehlke and Biehl, 1934; Yule and Kendall, 1950; Blalock, 1964; Arbia, 1989), bivariate regression (Clark and Avery, 1976; Arbia, 1989), multivariate statistical analysis (Fotheringham and Wong, 1991), spatial interaction models (Openshaw, 1977; Batty and Sikdar; 1982; Putman and Chung, 1989; Amrhein and Flowerdew, 1992), and location-allocation modelling (Hillsman and Rhoda, 1978; Goodchild, 1979; Bach, 1981; Casillas 1986; Fotheringham, Densham and Curtis, 1995). The findings from the aforementioned studies raise our skepticism on the reliability of the outcome of any spatial study relying on the use of areal data. Though the degree of the impact has been found to vary from study to study, this unpredictability further complicates the problem and stresses the need for more insight, and solutions, to the problem.

While relevant research effort has concentrated mostly on revealing the MAUP, the search for effective solutions has not been widely attempted, at least not with satisfactory results. According to Wong (1996), past attempts are categorized into three approaches: data manipulation, technique-oriented and error modeling. The data manipulation approach is based on the suspicion that the MAUP would vanish if the chosen areal units can be justified one way or another, instead of for administrative convenience (Openshaw 1977; Fotheringham and Wong 1991). Openshaw (1977, 1996), You, Nedovic-Budic and Kim (1997), Ding (1998) and Guo (2000) and Alvanides, Openshaw and Macgill (2001) are among those researchers who develop methods for creating optimal zones with respect to predefined objective functions. The technique-oriented approach, on the other hand, is based on the argument that the MAUP effect might have been a result of using inappropriate models or statistical techniques in analyzing aggregated spatial data (Amrhein and Flowerdew 1991, Tobler 1991). This leads to Tobler's (1991) proposal of abandoning the unsuitable classical statistical techniques and replacing them with frame independent analyses. Another group of researchers (e.g., Steel, Holt and Tranmer

1994) recognize that, when analysis moves from one spatial scale to another, relationships among variables and among spatial entities also change. Instead of searching for techniques immune to such scale effects, they adopt the error modeling approach of explicitly documenting variations derived from changing scale, and incorporating these changes into modeling and analysis.

To date, a general, workable solution to the problem does not yet exist and the MAUP remains one of the most stubborn problems in geography and spatial science (Wong 1996, Fotheringham, Brunsdon and Charltom 2000). However, not all attempts in resolving the problem have been futile. As Miller (1999) surveys recent work on the MAUP, he suggests "it is clear that antecedent factors can be controlled and [the problem's] effects predicted, particularly within specific application contexts" (p.375). That is, in order to reduce, or remove, the effects of MAUP, whether over temporal, spatial or other domains, it is necessary to know something about the general nature of that phenomenon. In the temporal instances, there are often strong organizing principles associated with the observations that give rise to self-similarity, which analysts can exploit to perform generalization. For example, it is intuitive that traffic volumes vary significantly for peak and off-peak hours. Based on this understanding, analysts therefore produce level-of-service measures for peak versus off-peak periods as opposed to some random temporal units. What often makes the spatial instances difficult is our lack of such intuition about the phenomenon at hand and analysts are thus required to decide on the spatial units before attempting to study the phenomenon.

In residential location choice models, using different zonal configurations may lead to parameter estimates as a manifestation of the MAUP. This is probably why past studies showed inconsistent or unintuitive findings about the effects of various choice determinants. Model parameter estimates derived from arbitrarily defined zones should be interpreted only with respect to these zones and do not correctly reflect residents' choice behavior unless the zones are

coterminous with neighborhoods as perceived by the residents. In order to reduce, or remove, the effects of MAUP in the study of residential location choice, analysts need a more precise and behavioral-oriented definition of neighborhood for practical measurement of neighborhood factors, improved conceptual understanding, as well as better transferability of models.

3.2.2 What is a Neighborhood?

Neighborhood is a vague, difficult-to-define concept. To the best of my knowledge, previous empirical studies of residential location choice have never formally addressed the issue of neighborhood definition. Scholars investigating the significance of neighborhood for individuals' behavior and well-being often do not rise to the challenge of providing the term with an explicit definition. As Galster (2001, p.2111) puts it, "[u]rban social scientists have treated 'neighborhood' in much the same way as courts of law have treated pornography: a term that is hard to define precisely, but everyone knows it when they see it". When spatial definition of neighborhood is required for the purpose of quantitative analysis, "most social scientists and virtually all studies of neighborhoods … rely on geographic boundaries defined by the Census Bureau or other administrative agencies… [which] offer imperfect operational definitions of neighborhoods for research and policy" (Sampson *et al*, 2002, p.445). This widespread practice suggests that perhaps we don't really know 'it' − at least not as well as we think − when we see 'it'. In order to better understand the nature of neighborhood, a collection of approaches in defining the term is reviewed and discussed below. The review is by no means exhaustive as the focus is on definitions that will help formulate operational units for neighborhoods (see Galster (2001) for a more extensive survey of the literature).

An area in which neighborhood definition plays an important role is the study of neighborhood effects, which refers to the neighborhood influences on the well-being and behavior of families, and often children in particular. One of the pioneers in the field, Park

(1915), points out that cities are generally outlined by their physical geography, natural advantages, and transportation systems. The processes of human nature then proceed to shape cities through competitions for efficient locations among businesses and individuals. These informal processes result in the formation of neighborhoods – naturally segregated localities that share similar sentiments, traditions and history. Followers of Park's school of thought tend to consider neighborhoods as discrete, non-overlapping communities, leading to the common use of administratively defined zones for analyzing neighborhood effects.

Later, Suttles (1972) argues that, in addition to being the result of free-market competition, some communities' identities and boundaries are imposed by outsiders. Suttles also suggests that neighborhoods are best thought of not as distinct areas of a city, but rather as a hierarchy of ecological grouping at four levels. At the lowest level is the 'local networks and the face-block', namely, a grouping of residents who share the same local facilities and residential condition because of their proximity to each other. A 'neighborhood', defined at this level, is usually different for each person and is unlikely to have any sharp boundaries. The second level is labeled the 'defending neighborhood', defined as "the smallest area which possesses a corporate identity known to both its members and outsiders" (p. 57). Its size may vary, but generally large enough to include a complement of establishments (grocery, liquor store, church, etc.) which people use in their daily round of local movements. The next level, the 'community of limited liability', is typically a construct imposed by external commercial and governmental interests, which could be political, educational or religious. Residents may be associated with multiple communities whose boundaries are fragmented and overlapping. The highest level in the neighborhood hierarchy is the 'expanded community of limited liability'. These are large scale community organizations referring to entire sectors of a city, such as North Austin, whose identity usually arise from government policies and programs.

Galster (2001) defines neighborhood as a 'complex commodity' that is produced by the same actors – households, businesses, property owners and local governments – that consume them. Neighborhood is a bundle of spatially based attributes, including structural, infrastructural, demographic, class status, tax/public service package, environmental, proximity, political, social-interactive, and sentimental characteristics. Consistent with Suttle's (1972) multi-scale view of neighborhood, Galster argues that the geographical scale across which a neighborhood attribute varies is often different for different attributes. Consumers' perceived delineation of a neighborhood thus depends on the particular neighborhood attributes of interest. This view is also shared by O'Campo (2003), who contends that the processes operating in the neighborhood environment are often many and that the ideal geographic units of analysis for different social processes may not be compatible.

The multi-scale structure of neighborhood can also be viewed as residents having multiple neighborhood memberships. As different processes (social, educational or religious) unfold, a household can identify its local identity through its residential neighbors, the school the children go to, its membership in a church, etc. These group memberships lead to spatial clusters, some of which may be objectively recognizable (such as school catchment area, red-light district, or gated community). In other cases, however, there are often no clear cutoff points for determining how far social contact or other processes reach. The boundaries for such neighborhood attributes are subjective and fuzzy. As group memberships of individuals evolve with changes in their role in the network of social interaction and their stage in life cycle, their perceptions of neighborhood also change (Horton and Reynolds 1971). The perception may also be influenced by race (Lee et al 1991) and gender (Guest and Lee 1984). Furthermore, an individual's perceived neighborhood also depends on where she or he lives: "an individual living on the boundary of a census tract probably has more in common with residents of the adjoining

tract than with residents on the far side of his own" (Dubin 1992, p. 435). The concept that no set of fixed neighborhood boundaries can accurately describe an urban area is referred to as 'sliding neighborhoods'.

Motivated by the uncertainty about how to construct operational units for neighborhoods in view of the many factors influencing residents' perception, Coulton et al. (2001) set out to examine the residents' perception through their mental maps. They asked 140 parents of minor children from the City of Cleveland to draw a map of what they considered as the boundaries of their neighborhoods. The study found evident discrepancies between resident-defined neighborhoods and census geography. Mental maps of neighborhoods typically include portions of at least two census tracts and three block groups, even though the average size is close to the size of a census tract. The study also demonstrated that individuals residing in close proximity and homogenous on race, age and gender can differ markedly from one another in how they define the physical space of their neighborhood. This variability renders the task of defining resident-perceived neighborhoods a very challenging proposition. The authors conclude by suggesting further research on mental maps of neighborhoods. However, Shinn and Toohey (2003) argue that even residents' hand drawn mental maps, which may be influenced by neighborhood names or generally acknowledged definitions, may not reflect the geographic areas that truly affect them. Areas where residents spend time through which they often travel en route to pursue activities may be more influential. The size of 'functional neighborhood' may "vary systematically with a person's age, health, or employment status" (p. 449).

Grannis (1998, 2003) also attempts to construct a practical representation of neighborhoods. He contends that street networks are one of the primary tools populations use to organize themselves in urban settings and that "the network of tertiary [small, residential-type] streets give rise to a network of neighborly relations" (1998, p.1560). He argues that pedestrian

streets give rise to close-knit communities where residents consider the boundaries between house and street space to be quite permeable and that networks of neighborly relations will emerge from and reflect networks of pedestrian streets. In a subsequent effort, Grannis (2003) models cities as multiple independent 'islands' – discontinuous networks of pedestrian streets that are separated by major thoroughfare. By comparing these islands with residents' cognitive maps of their neighborhood, he shows that, while islands circumscribe residents' perception of their neighborhoods, residents typically perceive only a portion of their island as their neighborhood. Like Coulton et al. (2001), he is unable to construct operational spatial units as close proxies for perceived neighborhoods.

The studies discussed above reflect the well-recognized difficulty in defining neighborhood, both at the conceptual and the operational levels. While the question of neighborhood definition remains to be further explored, the existing literature sums up to a few pointers for constructing neighborhood boundaries. First, administratively defined units do not represent real neighborhoods and are thus imperfect operational definitions of neighborhoods for research and policy. However, census geography in terms of tracts, block groups and blocks are reasonably consistent with the notion of neighborhoods as nested ecological structures. Second, an urban region can be viewed as a mix of fixed (objectively recognizable boundaries such as major roads, geographical barriers and political demarcations) and sliding (subjective boundaries that depend on the characteristics and location of the residents) neighborhoods. Third, and as a result of the previous two points, the size of a perceived neighborhood can range from the size of multiple census blocks to multiple tracts. Lastly, a neighborhood has a geographical reference, but its meaning depends on function and domain. The relevant units depend on the specific process, or the outcome of the process, being studied.

3.3 Summary

This chapter has identified why proportional substitutability among choice alternatives is a behaviorally unrealistic assumption. Of the alternative modeling approaches reviewed, the GEV and the MMNL formulations show the most promise for accommodating flexible substitution patterns, though the flexibility is usually gained at the expense of computational difficulty unless the models are 'cleverly' customized for the application context.

The chapter has also examined the problems arisen from grouping elemental alternatives into, and constructing spatially aggregate measures over, arbitrarily or administratively defined zones, as opposed to neighborhood definitions as perceived by decision makers. The experience from past research efforts aimed at conceptualizing the nature of neighborhood suggests that neighborhood is intrinsically hierarchical and is continuously shaped by the infrastructure and the many ecological and social processes that take place in the urban environment. The hierarchical organization and the spatial boundaries of neighborhood are very much domain dependent. In certain contexts they can be described by objectively recognizable spatial delineations while, in other situations, they are constructed by individuals' perception, which may be influenced by race, gender, age, stage in life cycle, social contact and physical location. This dynamic nature of neighborhood renders the grouped alternatives model methodologically flawed. By not appropriately considering neighborhood attributes over the area that really matters to the decision makers, many of the studies reviewed in Section 2.6 were likely to have produced biased parameter estimates that lead to erroneous interpretations. A more behaviorally approach would be to incorporate in a single model structure neighborhood attributes measured based on a hierarchy of spatial definitions, which represent either fixed or sliding neighborhoods depending on the nature of the attribute.

CHAPTER 4

ADDRESSING INTERALTERNATIVE CORRELATIONS

When no micro-level data is available about the locality and characteristics of individual dwellings, it is inevitable for analysts to consider groupings of dwelling units as choice alternatives. Often, the groupings are constructed based on the spatial proximity of dwellings so that predefined zones that constitute a given study region are considered as the alternatives. Models then assume the MNL structure. For reasons explained in Section 3.1.1, the presence of correlation among the unobserved utilities of neighboring alternatives renders the proportionate substitutability assumption embedded in such MNL-based, grouped alternatives models inappropriate.

In searching for a model structure to accommodate the inter-alternative spatial correlation, a number of advanced discrete choice models have been examined earlier in Section 3.1.2 and the MMNL and the GEV class of models were identified as allowing very flexible sustainability. In particular, the GNL model is conceptually appealing in that it allows the fractional allocation of each alternative to a set of nests, each of which having a distinct dissimilarity parameter. Its closed-form formulation is also a computational advantage. The MMNL model, on the other hand, is especially useful for capturing the correlations among the unobserved of utility over alternatives due to decision maker's differential tastes.

The aim of this chapter is to develop, based on the GNL and the MMNL model structures, a modeling framework that accommodates the unobserved correlation among choice alternatives as well as the unobserved heterogeneity across individuals. The chapter begins with Section 4.1, which describes a version of the GNL model first proposed by Wen and Koppelman's (2001) to

work with nests of paired alternatives. Section 4.2 describes how this paired nests structure can be customized for spatial contexts. The spatially correlated logit (SCL) model developed based on the paired nested structure is formally stated in Section 4.3. Combining the SCL with the MMNL formulation, Section 4.4 derives the mixed spatially correlated logit model (MSCL) which is capable of accommodating heterogeneity across individuals in their responsiveness to exogenous determinants of residential location choice. The empirical application of the MSCL model to the Dallas-Fort Worth region in Texas is described in Section 4.6, following by a summary of the chapter in Section 4.7.

4.1 Paired Generalized Nested Logit Model

As part of their effort to prove that the PCL model is a restricted version of the GNL model, Wen and Koppelman (2001) described a special case of the GNL that includes one nest for each pair of alternatives, as in the PCL model. This paired GNL (PGNL) model has the form

$$P_n = \sum_{j \neq i} \left(\frac{(\alpha_{i,ij} e^{V_{n,i}})^{1/\mu_{ij}}}{(\alpha_{i,ij} e^{V_{n,i}})^{1/\mu_{ij}} + (\alpha_{j,ij} e^{V_{n,j}})^{1/\mu_{ij}}} \times \frac{\left((\alpha_{i,ij} e^{V_{n,i}})^{1/\mu_{ij}} + (\alpha_{j,ij} e^{V_{n,j}})^{1/\mu_{ij}}\right)^{\mu_{ij}}}{\sum_{k=1}^{I-1} \sum_{l=k+1}^{I} \left((\alpha_{k,kl} e^{V_{n,k}})^{1/\mu_{kl}} + (\alpha_{l,kl} e^{V_{n,l}})^{1/\mu_{kl}}\right)^{\mu_{kl}}} \right) \quad (4.1)$$

If the allocation parameters, $\alpha_{i,ij}$, are equal for all paired nests, then the PGNL model reduces to the PCL model. The non-equal allocation to nests in the PGNL model allows greater freedom in the magnitude of cross-elasticity than is allowed by the corresponding PCL model. Also an important feature is that the PGNL formulation allows an allocation of zero for an alternative to a nest and the elimination of nests for which both alternatives have zero allocation.

4.2 Paired Nested Structure

A common approach in the spatial analysis literature for capturing spatial correlation is to allow immediately adjacent observations to share common unobserved characteristics. This

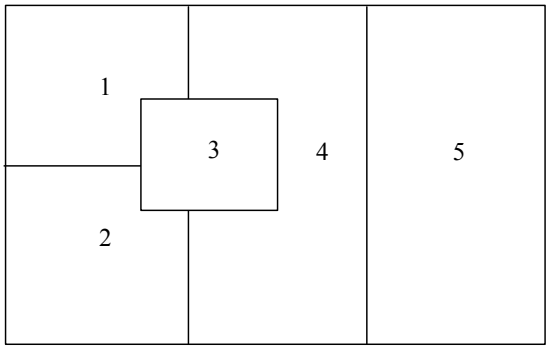
dissertation research adopts a similar approach to capture spatial correlation for the aggregate residential choice problem. This is achieved by defining a two-level, paired nested structure, for which the PGNL model can be easily applied. The structure consists of as many nests as the number of adjacent zone pairs. This paired nested stricture is formally defined as follows.

First, denote the alternative residential zones by i, $i = 1\ldots, I$. Let ω_{ij} be a dummy variable that takes a value of 1 if zone j is adjacent to zone i and 0 otherwise (by convention, $\omega_{ii} = 0$). The number of zones adjacent to zone i is therefore $\sum_{j=1}^{I} \omega_{ij}$. Next, define a two-level nested structure, in which each pair of adjacent zones (i, j) is represented by a nest and the total number of paired nests is $\sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \omega_{ij}$. As a zone can have more than one neighboring zone, each zone i is allocated to the paired nest (i, j) based on an allocation parameter $\alpha_{i,ij}$ such that $\sum_{j} \alpha_{i,ij} = 1$. That is, the total allocation of zone i across all pairings of i with other alternatives is unity. One way of defining $\alpha_{i,ij}$ is to assume that the sensitivity to changes in neighboring spatial units is larger for a zone with fewer neighboring zones and that zone i is equally correlated with each neighboring zone. Accordingly, the allocation parameter for zone i is defined as:

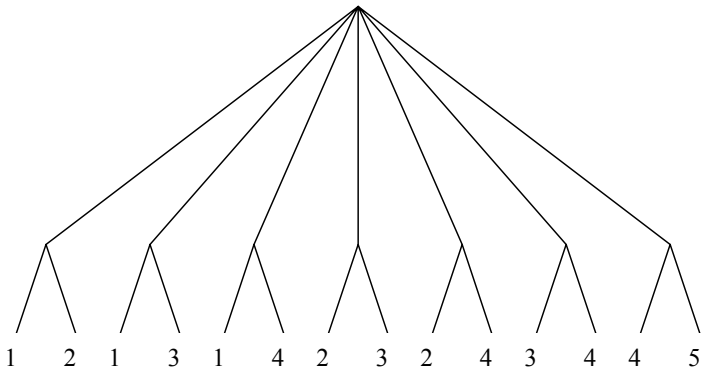
$$\alpha_{i,ij} = \frac{\omega_{ij}}{\sum_{k} \omega_{ik}}. \tag{4.2}$$

An example is illustrated in Figure 4.1 to help clarify the concepts and notations introduced so far. For the five spatial units configured as shown in Figure 4.1(a), the corresponding paired nested structure contains seven nests of zone pairs as depicted in Figure 4.1(b). The corresponding contiguity matrix that defines the ω_{ij}'s and the resulting allocation parameters $\alpha_{i,ij}$ are provided in the table in Figure 4.1(c).

(a) Spatial configuration



(b) Paired nested structure for generating spatial correlation between adjacent spatial units



(c) Contiguity (Allocation) Matrix

Alternative	Spatial Units					Number of Adjacent Spatial Units
	1	2	3	4	5	
1	0 (0)	1 (0.33)	1 (0.33)	1 (0.33)	0	3
2	1 (0.33)	0 (0)	1 (0.33)	1 (0.33)	0	3
3	1 (0.33)	1 (0.33)	0 (0)	1 (0.33)	0	3
4	1 (0.25)	1 (0.25)	1 (0.25)	0 (0)	1 (0.25)	4
5	0	0	0	1 (1)	0 (0)	1

Figure 4.1 A simple example of residential choice among five spatial units

4.3 Spatially Correlated Logit Model

Once the residential zones are translated into the paired nested structure in the manner described in the preceding section, the application of the PGNL formulation seems straightforward. However, the direct application of Equation (4.1) would result in as many dissimilarity parameters to estimate as the number of zone pairings. This is undesirable for most practical situations where the number of zones is large. Instead, the SCL model is constructed based on the assumption that the dissimilarity parameter is the same for all pairings of residential zones.

Specifically, the SCL model is derived based on the following function:

$$G\left(e^{V_{n,1}},\ldots,e^{V_{n,I}}\right)=\sum_{i=1}^{I-1}\sum_{j=i+1}^{I}\left[(\alpha_{i,ij}e^{V_{n,i}})^{1/\mu}+(\alpha_{j,ij}e^{V_{n,j}})^{1/\mu}\right]^{\mu}, \tag{4.2}$$

where the double summation includes all pairs of alternatives in the choice set, $\alpha_{i,ij}$ is as defined earlier, $V_{n,i}$ represents the deterministic component associated with zone i, and μ is the dissimilarity parameter capturing the correlation between adjacent zones, $0<\mu\le 1$. Substituting Equation (4.2) into Equation (1.10) yields the multivariate extreme value distribution:

$$F(\varepsilon_{n,1},\varepsilon_{n,2},\ldots,\varepsilon_{n,I})=\exp\left\{-\sum_{i=1}^{I-1}\sum_{j=i+1}^{I}\left[(\alpha_{i,ij}e^{-\varepsilon_{n,i}})^{1/\mu}+(\alpha_{j,ij}e^{-\varepsilon_{n,j}})^{1/\mu}\right]^{\mu}\right\}, \tag{4.3}$$

where $\varepsilon_{n,i}$ denotes the stochastic utility for zone i. Substituting Equation (4.2) into Equation (3.2) gives the probability of decision maker n choosing zone i:

$$P_{n,i}=\frac{\sum_{j\neq i}(\alpha_{i,ij}e^{V_{n,i}})^{1/\mu}\left((\alpha_{i,ij}e^{V_{n,i}})^{1/\mu}+(\alpha_{j,ij}e^{V_{n,j}})^{1/\mu}\right)^{\mu-1}}{\sum_{k=1}^{I-1}\sum_{l=k+1}^{I}\left((\alpha_{k,kl}e^{V_{n,k}})^{1/\mu}+(\alpha_{l,kl}e^{V_{n,l}})^{1/\mu}\right)^{\mu}} \tag{4.4}$$

The probability function can be rewritten as

$$P_{n.i} = \sum_{j \neq i} \frac{(\alpha_{i,ij} e^{V_{n,i}})^{1/\mu}}{(\alpha_{i,ij} e^{V_{n,i}})^{1/\mu} + (\alpha_{j,ij} e^{V_{n,j}})^{1/\mu}} \times \frac{\left((\alpha_{i,ij} e^{V_{n,i}})^{1/\mu} + (\alpha_{j,ij} e^{V_{n,j}})^{1/\mu}\right)^{\mu}}{\sum_{k=1}^{I-1} \sum_{l=k+1}^{I} \left((\alpha_{k,kl} e^{V_{n,k}})^{1/\mu} + (\alpha_{l,kl} e^{V_{n,l}})^{1/\mu}\right)^{\mu}}$$

$$= \sum_{k=1}^{K} P_{i|k} P_k \ . \qquad (4.5)$$

The above expression can be derived directly from restricting $\mu_{ij} = \mu, \forall i, j$ in Equation (4.1). Similarly, the self- and cross-elasticities for the SCL model can be derived from those for the GNL model by letting $\mu_k = \mu, \forall k$ in Equation (3.12) and (3.13). The resulting expressions for the elasticities for the SCL model are provided in the second row of Table 4.1, while those for the MNL model are shown in first row of the same table. Note that the index n is dropped from the expressions for simplicity. The table provides a clear contrast between the two models. The cross-elasticity for the MNL model reflects the IIA property (equal cross-elasticities of the effect of alternative i on any alternative j). The cross-elasticity expression in the SCL model, on the other hand, indicates equal proportionate change in all non-adjacent alternatives to i due to a change in the utility of alternative i. However, the cross-elasticities are higher for spatial units contiguous to i.

Table 4.1 Expressions for the direct and cross-elasticities in the MNL, SCL and MSCL Models

Model	Direct elasticity[1]	Cross-elasticity[2]
MNL	$(1-P_i)\beta_m x_{im}$	$-P_i\beta_m x_{im}$
SCL	$\left\{\sum_{j\neq i} P_{i\|ij}P_{ij}\left[(1-P_i)+\left(\frac{1}{\mu}-1\right)(1-P_{i\|ij})\right]\right\}\frac{\beta_m x_{im}}{P_i}$	$-\left[P_i+\frac{\left(\frac{1}{\mu}-1\right)P_{i\|ij}P_{ij}P_{j\|ij}}{P_j}\right]\beta_m x_{im}$ if i and j are spatially contiguous $-P_i\beta_m x_{im}$ if i and j are not contiguous
MSCL	$\left\{\int_{\beta=-\infty}^{\infty}\left[\sum_{j\neq i}(R_{ij}\|\beta)\beta_m\right]f(\beta\|\theta)d\beta\right\}\frac{x_{im}}{P_i}$, where $R_{ij}\|\beta=(P_{i\|ij}\|\beta)(P_{ij}\|\beta)\left[(1-P_i\|\beta)+\left(\frac{1}{\mu}-1\right)(1-P_{i\|ij}\|\beta)\right]$	$\left\{-\int_{\beta=-\infty}^{\infty}\left[(P_i\|\beta)(P_j\|\beta)+\left(\frac{1}{\mu}-1\right)(P_{i\|ij}\|\beta)(P_{ij}\|\beta)(P_{j\|ij}\|\beta)\right]\beta_m f(\beta\|\theta)d\beta\right\}\frac{x_{im}}{P_j}$ if i and j are spatially contiguous $\left\{-\int_{\beta=-\infty}^{\infty}(P_i\|\beta)(P_j\|\beta)\beta_m f(\beta\|\theta)d\beta\right\}\frac{x_{im}}{P_j}$ if i and j are not contiguous

[1] Direct elasticity refers to the percentage change in the choice probability of alternative i due to a 1% change in the m^{th} variable associated with alternative i.

[2] Cross-elasticity is the percentage change in the choice probability of alternative j due to a 1% change in the m^{th} variable associated with alternative i.

4.4 Mixed Spatially Correlated Logit Model

In residential location choice as well as other spatial choice contexts, individuals' responsiveness to exogenous determinants will likely vary across individuals, due to both observed and unobserved taste preferences. As of the MNL model, the SCL model described in the preceding section maintains the assumption of homogenous responsiveness. In order to account for observed taste variations, interaction terms can be introduced into the utility function in the same way as described in Section 2.2.1. For example, to allow for different sensitivity to commute time between man and female workers in the residential location choice model, one adds an interaction variable for commute time and women (dummy variable). However, the SCL model can not accommodate any variation in responsiveness to commute time due to unobserved factors, such as commuters' experience and attitude.

This research adopts the mixing structure, similar to that described in Section 3.1.2.2 for the mixed MNL models, to accommodate unobserved response heterogeneity. That is, the coefficient vector β embedded in the V_i vector in the SCL model is assumed to be multivariate normal with a vector θ of underlying moment parameters. Let f represent the density function of the multivariate distribution and let F be the corresponding distribution function. Then, the choice probability of alternative i in the mixed SCL (MSCL) model may be written as:

$$P_{n,i} = \int_{-\infty}^{\infty} (P_{n,i} \mid \beta) f(\beta \mid \theta) d\beta, \tag{4.6}$$

where

$$P_{n,i} \mid \beta = \frac{\sum_{j \neq i} (\alpha_{ij} e^{\beta' x_{n,i}})^{1/\mu} \left[(\alpha_{i,ij} e^{\beta' x_{n,i}})^{1/\mu} + (\alpha_{j,ij} e^{\beta' x_{n,j}})^{1/\mu} \right]^{\mu-1}}{\sum_{k=1}^{I-1} \sum_{l=k+1}^{I} \left[(\alpha_{k,kl} e^{\beta' x_{n,k}})^{1/\mu} + (\alpha_{l,kl} e^{\beta' x_{n,l}})^{1/\mu} \right]^{\mu}}. \tag{4.7}$$

The resulting self- and cross-elasticities for the MSCL model are provided in the last row of Table 4.1. The cross-elasticity expressions in the MSCL model do not exhibit the equal proportional change propensity for non-contiguous alternatives to i. This is different from the cross-elasticity expressions for the SCL model.

As pointed out in Section 3.1.2.2, the MMNL structure could have ben used for accommodating within a single framework both the disproportionate substitutability, and hence spatial correlation, among alternatives as well as any unobserved response heterogeneity. However, in so doing, the dimensionality of the integral in the choice probability would be equal to the number of random elements in the vector β plus the number of paired nests of adjacent alternatives. The MSCL model is able to accommodate the same effect with the dimensionality of the integral equal to the number of random elements in the coefficient vector β. In most practical contexts, the additional dimensionality incurred from using the MMNL approach would result in great computational difficulties.

4.5 Model Estimation

The parameters to be estimated in the MSCL model include the scalar μ representing spatial correlation and the θ vector characterizing the multivariate normal distribution of the β parameters. For the discussion below, μ is absorbed in the parameter vector θ for the ease of presentation.

The estimation of the MSCL model can be pursued using the maximum likelihood principle described in Section 2.3. The evaluation of the log-likelihood function defined in Equation (2.19) requires the evaluation of the multi-dimensional integrals in the choice probability expression (Equation (4.6)). This can be achieved by applying a simulation method to approximate the multi-dimensional integrals and maximize the resulting simulated log-likelihood

function. The simulation technique entails computing the choice probability in Equation (4.6) at several values of the β vector drawn from the multivariate normal distribution for a given value of the parameter vector θ and averaging the integrand values. Formally, let $\widetilde{P}_{n,i}^{m}(\theta)$ be the realization of the choice probability for the n th household in the m th draw ($m = 1,2,\ldots,M$). The choice probabilities are then computed as:

$$\widetilde{P}_{n,i}(\theta) = \frac{1}{N}\sum_{m=1}^{M}\widetilde{P}_{n,i}^{m}(\theta) \tag{4.8}$$

where $\widetilde{P}_{n,i}(\theta)$ is the simulated choice probability of the n th household choosing alternative i given the parameter vector θ. $\widetilde{P}_{n,i}(\theta)$ is an unbiased estimator of the actual probability and its variance decreases as the number of draws, M, increases. It also has the appealing properties of being smooth (*i.e.*, twice differentiable) and being positive for any realization of the finite M draws.

The simulated log-likelihood function is constructed by substituting Equation (4.8) into Equation (2.19):

$$SLL(\theta) = \sum_{n=1}^{N}\sum_{j=1}^{J}\delta_{n,j}\ln\widetilde{P}_{n,j}(\theta). \tag{4.9}$$

The parameter vector θ is estimated as the vector value that maximizes the above simulated function. Under rather weak regularity conditions, the maximum simulated log-likelihood estimator is consistent, asymptotically efficient, and asymptotically normal (see Hajivassiliou and Ruud, 1994; Lee, 1992).

In computing the simulated probability, $\widetilde{P}_{n,i}(\theta)$, if the integrand is computed at a sequence of random, or pseudo-random, points, the method is known as the Pseudo Monte Carlo (PMC) simulation method. The method suffers from the drawback of slow asymptotic convergence rate and thus requires a high number of simulation draws. An alternative approach

is the quasi-Monte Carlo (QMC) method, which involves the evaluation of the integrand at "cleverly" crafted non-random and more uniformly distributed points within the domain of integration (Bhat, 2001). The sequences of evaluation points used in the QMC method are generally referred as the quasi-random sequences. Of the several quasi-random sequences that may be employed for the simulation, the Halton sequence is chosen for the empirical analysis presented in the subsequent section because of its conceptual simplicity. The Halton simulation method has been shown by Bhat (2002b), Hensher (1999) and Train (1999) to out-perform the traditional PMC methods for estimating mixed discrete choice models.

Details of the Halton sequence and the procedure to generate this sequence are available in Bhat (2002b). In short, the Halton sequence is generated by choosing a prime number r ($r \geq 2$) and expanding the sequence of integers $0,1,2,\ldots g,\ldots,G$ such that

$$g = \sum_{l=0}^{L} b_l r^l \quad , \tag{4.10}$$

where $0 \leq b_l \leq r-1$ and $r^L \leq g \leq r^{L+1}$. Thus, $g\left(g = 1,2,\ldots,G\right)$ can be represented by the r*-adic* integer string $b_L b_{L-1} \ldots b_1 b_0$. The Halton sequence in prime base r is then obtained by taking the radical inverse of $g\left(g = 1,2,\ldots,G\right)$ by reflecting through the radical point:

$$\varphi_{(r)}\left(g\right) = 0.b_0 b_1 \ldots b_{L-1} b_L \text{ (in base } r) = \sum_{l=0}^{L} b_l r^{-l-1} \ . \tag{4.11}$$

The sequence above is very uniformly distributed in the interval (0,1) for each prime number r. The Halton sequence for evaluating a K-dimensional integral is obtained by pairing K one-dimensional sequences based on K relatively prime integers.

4.6 Empirical Application

An empirical application of the MSCL model has been conducted for the cities of University Park, Highland Park, and Dallas which are situated in North-Central Texas. As shown

in Figure 4.2, the three cities form part of the Dallas County and represent 98 out of the 383 Transport Analysis and Processing (TAP) zones in the county.

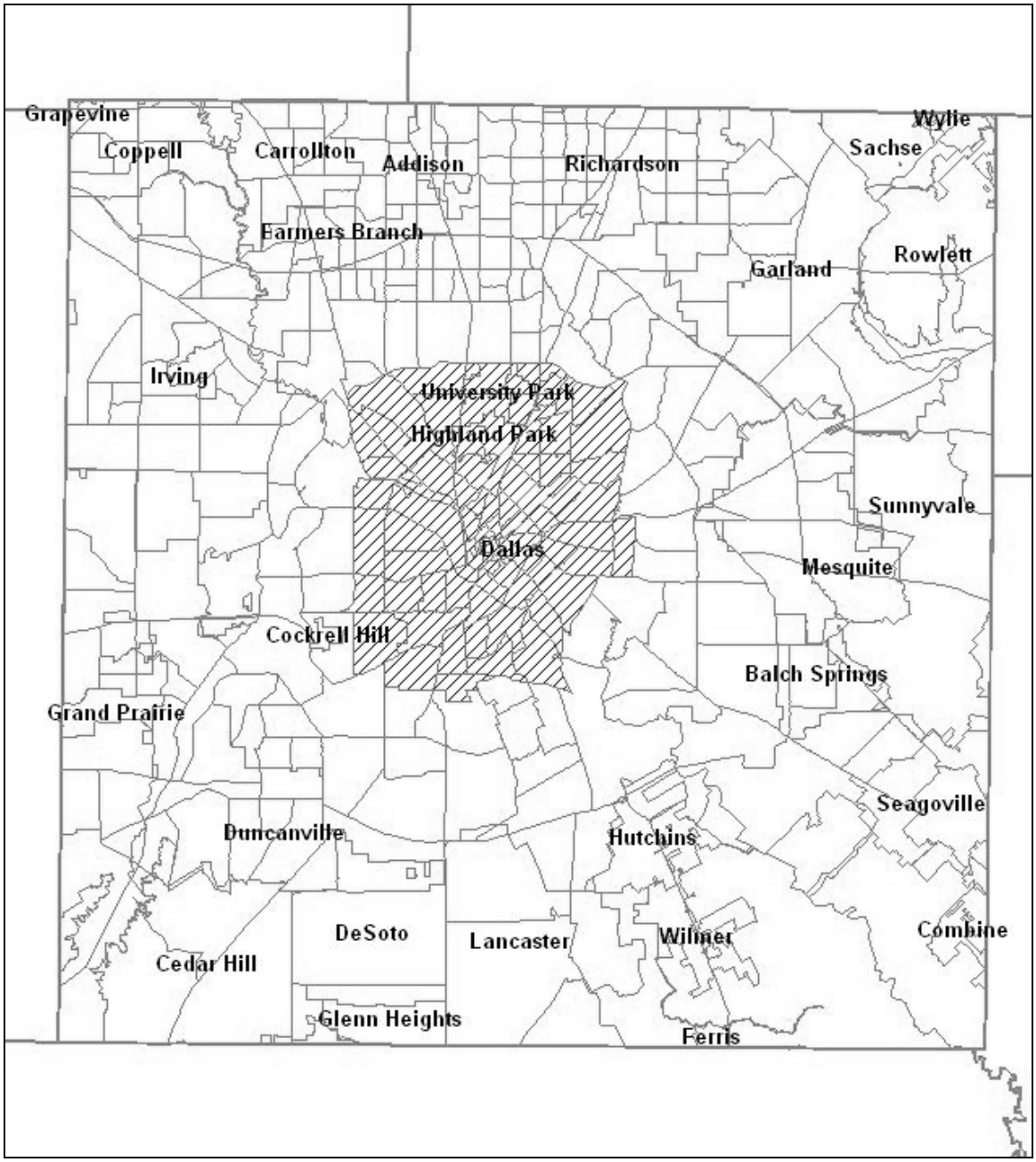


Figure 4.2 The study region (shaded area) includes three cities in North-Central Texas.

4.6.1 Data Source and Sample

The primary source of data is the 1996 Dallas-Fort Worth (D-FW) metropolitan area household activity survey. This survey collected information about travel and non-travel activities undertaken during a weekday by members of 4839 households, as well as the residential zones of households. The survey also obtained individual and household socio-demographic information. In addition to the activity survey, four other data sets associated with the D-FW metropolitan area were used: land-use/demographic coverage data, census data, zone-to-zone travel level-of-service (LOS) data, and school-rating data. Both the LOS and the land-use data files were obtained from the North Central Texas Council of Governments (NCTCOG), which is a voluntary association of local governments from 16 counties centering from two urban centers of Dallas and Fort Worth. The LOS file provides information on travel between each pair of the 919 TAP zones in the North Central Texas region. The file contains the inter-zonal distances as well as peak and off-peak travel times and costs for transit and highway modes. The land-use coverage file contains acreage by land-use purposes (including water area, park land, roadway, office, retail and etc.) for each of 5938 traffic survey zones (TSZ) in the region. The census data provides additional socio-demographic information such as the distribution of household ethnicity and housing cost in the zones. Data about school ratings is compiled in-house from the 1996 district summary of the Accountability Rating System (ARS) for Texas Public Schools and School Districts. The ARS is released on a yearly basis by the Texas Education Agency. The schools are classified into 4 levels: exemplary, recognized, acceptable and low performing (or unacceptable). The criteria for ranking are summarized in Table 4.2.

The empirical application focuses on households with one and only one worker. 236 of such households were found and extracted from the survey data to develop the sample for model

estimation. The final sample contains for each household the 98 TAP zones that constitute the study area as choice alternatives.

Table 4.2 School quality ranking system used by the Texas Education Agency

School Ranking	Dropout Rate	Attendance Rate	Percent of Students Passing TAAS
Exemplary	1% or less	At least 94%	At least 90%
Recognizable	3.5% or less	At least 94%	At least 80%
Acceptable	6% or less	At least 94%	At least 40%
Low Performance	More than 6%	Less than 94%	Less than 40%

4.6.2 Variable Specifications

The five data sources identified in the previous section provide a rich set of variables to describe the utility associated with each of the alternative residential zones. The variables include various measures of the zones as well as interactions of socio-demographic characteristics of households with these zonal measures. The variables are categorized into six groups and are discussed below.

4.6.2.1 Size measures

Three size measures are considered: log of zonal area in squared mile, log of zonal population, and log of number of households in zone. They are used to correct for the grouping of elemental alternatives such that a large zone would have a higher probability of being selected than a small zone.

4.6.2.2 Commute-related variables

After the workers' employment zones are identified from the travel survey data, the auto commute time and distance from the employment zone to each alternative residential zone are extracted from the level-of-service data. Interactions of the commute time variable with the sex

and the race of the worker in each household are constructed to test for the presence of gender disparity in commute patterns and of greater spatial mismatch for minorities compared to non-minorities.

4.6.2.3 School quality measures

The school quality at each zone is represented by four dummy variables, each corresponding to one of the four rankings: exemplary, recognized, acceptable and low performing.

4.6.2.4 Socioeconomic and demographic variables

Three socioeconomic and demographic sets of variables are computed to test for the presence of residential segregation. The first set, the racial composition variables, are constructed as the zonal percentages of population belonging to each racial group. These zonal percentages are then interacted with dummy variables indicating the racial group of the sampled households. The second set of variables is computed by the absolute difference between household income and zonal median income as a measure of household income homogeneity. Similarly, the third set of variables, indicating household size homogeneity, are computed by the absolute difference between household size and zonal average household size,.

4.6.2.5 Land-use variables

These include the density measures and land-use composition measures. The density measures are computed by dividing the total number of households or people in a zone by the zone size (in acre). The land-use composition measures are computed by normalizing the acreage in different types of land uses by the total zonal acreage. The land use types considered include:

lakes and water, single-family housing, multi-family housing, industrial, offices, retail and services, institutions (schools, churches, etc.), and infrastructure.

4.6.2.6 Regional accessibility variables

A residential zone's attractiveness depends not only on the elements within the zone itself, but also how the zone relates spatially to the rest of the urban area. This is the motivation for considering regional accessibility measures for recreational, shopping, and employment activities. The Hansen-type (Fotheringham, 1986) accessibility measures are used. Large values of the accessibility measures indicate more opportunities for activities in close proximity of that zone, while small values indicate zones that are spatially isolated from such opportunities. The measures are defined as follows:

$$A^{\text{Rec}}{}_i = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{(\text{Park Land Acreage}_j)^{\gamma_{\text{Rec}}}}{(Impedance_{ij})^{\beta_{\text{Rec}}}}\right),$$

$$A^{\text{Ret}}{}_i = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{(\text{Number Of Retail Employment}_j)^{\gamma_{\text{Ret}}}}{(Impedance_{ij})^{\beta_{\text{Ret}}}}\right), \text{ and}$$

$$A^{\text{Emp}}{}_i = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{(\text{Number Of Basic Employment}_j)^{\gamma_{\text{Emp}}}}{(Impedance_{ij})^{\beta_{\text{Emp}}}}\right), \qquad (4.12)$$

where A^{Rec} represents the accessibility to recreational opportunities, A^{Ret} represents the accessibility to shopping opportunities, A^{Emp} represents the accessibility to basic employment opportunities, i is the zone index, and N is the total number of zones in the study region. $Impedance_{ij}$ is a composite highway auto impedance measure of travel between zone i and zone j. γ_{Rec}, β_{Rec}, γ_{Ret}, β_{Ret}, γ_{Emp}, and β_{Emp} are parameters that are estimated using a destination choice model of the form given below for the recreational activity purpose (similar formulations are used for the retail and basic employment categories):

$$V_{ij}^{rec} = \gamma^{rec} \times \ln\left(Park\ Land\ Acreage\right)_j - \beta^{rec} \times \ln\left(\text{Impedance}\right)_{ij}. \qquad (4.13)$$

where V_{ij}^{rec} is the utility presented by zone j for recreational participation to an individual in zone i. Assuming a MNL form for destination choice then leads to an accessibility index for zone i that is equal to $(1/N) \times \sum_j \exp\left(V_{ij}^{rec}\right)$. The functional form of V_{ij}^{rec} used in Equation (4.13) results in accessibility measures that are consistent with the formulations presented in Equation (4.12). The impedance expression used in the accessibility computations takes the form of a parallel conductance formula that accommodates multiple level-of-service measures and multiple modes (see Bhat et al, 1998 for a discussion of this formula). However, in the current empirical context, only highway auto level-of-service measures are used because of the lack of adequate transit observations in the destination choice model estimation. The highway auto impedance measure is in effective in-vehicle time units (in minutes) and is expressed as follows:

$$Impedance\ (\text{in IVTT min.}) = IVTT + \delta \times OVTT\ (\text{in min.}) + \eta \times COST\ (\text{in cents}). \qquad (4.14)$$

The estimated values of the δ, and η scalar parameters, and the γ and β vector parameters, are provided in Table 4.3. As can be observed, the only level of service variable that is relevant for recreational destination choice is in-vehicle time, while cost is not significant for employment destination choice. These results are perhaps a consequence of the strong multicollinearity in time and cost measures. For retail destination choice, the implied money value of time is \$6.05 per hour. The smaller estimated coefficient on out-of-vehicle time for the basic employment category suggests that, unlike in mode choice decisions, individuals place a smaller value on out-of-vehicle time than in-vehicle time when selecting employment destinations. This result may be a consequence of the dominance of in-vehicle time as the spatial separation measure when making destination choice decisions.

Table 4.3 Summary of destination choice model results for use in computing accessibility

Variable / Fit Measures	Purpose					
	Recreation		Retail		Basic Employment	
	Parameter	t-stat	Parameter	t-stat	Parameter	t-stat
Size measure	γ_{Rec} = 0.1376	8.92	γ_{Ret} = 0.2868	8.71	γ_{Emp} = 0.7554	61.40
Composite highway impedance	β_{Rec} = -2.6771	-40.92	β_{Ret} = -3.0779	-31.72	β_{Emp} = -2.6507	-86.15
In-vehicle time [1] (in mins.)	1.0000	--	1.0000	--	1.0000	--
Out-of-vehicle time (in mins.)	--	--	--	--	0.3385	8.13
Cost (in cents)	--	--	0.0992	2.5	--	--
Number of observations	1817		1206		4561	
Log-likelihood at convergence	-1912.60		-939.57		-4519.95	
Log-likelihood at equal shares	-3535.72		-2346.77		-8875.29	
Rho-squared value [2]	0.459		0.600		0.491	

1. Coefficient on this variable is constrained to one for identification purposes.
2. Computed as $1 - \frac{\text{log-likelihood of convergence}}{\text{log-likelihood at equal share}}$

4.6.3 Estimation Results

The data assembled for the 236 single-worker households are used to estimate two models structures, MNL and MSCL, for comparison. After a systematic process of eliminating variables found to be insignificant in earlier specifications and based on considerations of parsimony in representation, the best specifications obtained for the two models are presented in Table 4.4 and discussed below.

Table 4.4 Estimation Results of the MNL and MSCL Models

Variables	Multinomial Logit Model		Mixed Spatially Correlated Logit Model	
	Parameter	t-statistic	Parameter	t-statistic
Logarithm of zonal area (in mile2)	0.250	2.776	0.286	3.256
Commute time (in 100's of minutes)				
Mean	-3.673	-2.200	-4.409	-2.441
Standard Deviation[1]	0.000	—	6.504	1.180
Population density (in 10 persons/mile2)				
Mean	7.685	4.223	6.987	4.049
Standard Deviation[1]	0.000	—	9.358	1.600
Percentage zonal Hispanic population interacted with Hispanic dummy variable	1.235	1.214	1.094	1.127
Absolute difference between zonal median income and household income (in $100,000)	-1.270	-2.305	-1.056	-1.762
Percentage of zonal area occupied by multifamily housing				
Mean	-1.319	-2.063	-3.741	-2.919
Standard Deviation[1]	0.000	—	4.541	1.914
Work accessibility interacted with African-American household head dummy variable	-2.921	-3.891	-2.329	-3.310
Shopping accessibility	5.809	8.350	5.098	5.759
Dissimilarity parameter[2]	1.000	—	0.358	3.541
Number of observations	236		236	
Log-likelihood at convergence	-1013.43		-1000.93	

1. The standard deviations are implicitly constrained to 0 in the MNL model.
2. The dissimilarity parameter is implicitly constrained to 1 in the MNL model. The t-statistic for the dissimilarity parameter in the MSCL model is computed.

4.6.3.1 Results for the MNL model

The MNL model results are presented in the second main column of Table 4.4. The coefficient on the logarithm of zonal area has the expected positive sign, indicating that households are more likely to locate in larger zones than smaller zones. The effect of commute time has the expected negative sign. However, no presence of gender disparity or spatial mismatch for minorities is found. The coefficient on population density suggests that households are more likely to locate in zones with high population density. This may be due to better housing availability at these zones or merely a reflection of population clustering. The interaction effect of the percentage of Hispanic population with the dummy variable identifying if the head of the household is Hispanic indicates that Hispanic households tend to locate in zones with a high percentage of Hispanic population. Although this finding is in accordance with the racial segregation commonly observed in past studies, it is surprising that such an effect does not apply to African-American or Caucasian households. Residential segregation by income that is frequently sited in previous studies is also present in the current analysis. The only significant zonal land-use structure variable is the percentage of zonal area occupied by multifamily housing units. The parameter on this variable indicates a reluctance to locate in areas with a high percentage of multifamily units. Finally, the coefficients on the accessibility measures indicate that (a) after the effect of commute time is accounted for, African-American households are more likely to locate in areas with poor work accessibility, and (b) all households prefer locations that offer good accessibility to shopping.

The housing cost and school quality variables were, rather surprisingly, not statistically significant (see Sermons and Koppelman, 2001 for a similar result in their study of residential location in the San Francisco Bay Metropolitan area). The lack of influence of these two variables may be a consequence of the resolution used to represent their effect in the current analysis.

4.6.3.2 Results for the MSCL model

The MSCL model results in Table 4.4 are similar to those of the MNL model in terms of the directionality of the mean effect of variables. It is not possible to directly compare the magnitude of the effects of variables from the two models because of the difference in normalizations of the error term variances (the error term variance is normalized to 1 in the MNL model, but is normalized to a higher value in the MSCL model because of the presence of the random heterogeneity mixing terms). However, a couple of interesting observations may be drawn from the relative magnitudes of variable effects in the two models. First, commute time has a higher (mean) effect in the MSCL model relative to other variables, as can be observed from the higher ratio of the coefficient on commute time to other coefficients. Second, the mean negative effect of the percentage of zonal area occupied by multifamily households in a zone is also higher (relative to other variable effects) in the MSCL model. Clearly, the relative effects of variables are not the same in the two models.

Several variables in the MSCL model were specified to have random coefficients, but only those on commute travel time, zonal population density, and the percentage of zonal area in multifamily housing had some statistically significant impact. The results show that 75% of individuals like to live closer to their work place, while 25% prefer locations farther away from their work place. The random parameter on population density suggests that, while about 77% of households prefer zones with higher population density, 23% prefer zones with low population density. Similarly, 80% of households prefer zones with lower percentage of zonal area occupied by multifamily housing, while 20% prefer zones with a higher percentage. These results indicate heterogeneity in responsiveness across households.

Finally, the dissimilarity parameter of the MSCL model is much smaller than, and significantly different from, 1 (note that the t-statistic of the dissimilarity parameter in the table is

computed with respect to a value of 1). This result indicates a high level of spatial correlation in residential location choice, which the MNL model fails to recognize.

4.6.3.3 Elasticity effects and data fit

The MNL and MSCL models imply quite different patterns of inter-alternative competition. To demonstrate the differences, Table 4.5 presents the disaggregate self- and cross-elasticity values for a randomly selected individual in the sample (the table does not indicate an elasticity effect for accessibility to work because the randomly selected individual is Hispanic, and so the interaction effect of being an African-American and work accessibility does not apply). The numbers in the table indicate the self- and cross-elasticities due to an increase in the variables characterizing the actual chosen zone for the randomly selected individual (of course, any zone can be chosen for computing elasticity effects, but, due to space considerations and presentation ease, only the chosen zone is selected). The cross-elasticities are computed for a randomly selected zone that is adjacent to the zone whose attributes are changed, as well as for a randomly selected non-adjacent zone.

The MNL cross-elasticities are equal for each variable, reflecting the familiar independence from irrelevant alternatives (IIA) propensity. The cross-elasticities for the MSCL model are different due to (a) the correlation generated between each zone and its neighboring zones in the spatially correlated logit formulation, and (b) the random parameter specification on variables (the latter effect leads to different cross-elasticities even within the group of non-adjacent zones and the group of adjacent zones). Overall, the cross-elasticities of the MSCL model reflect the substantially higher sensitivity between adjacent zones (compared to non-adjacent zones) caused by spatial autocorrelation effects (note the substantially smaller values in the last column relative to the values in the last but one column).

Table 4.5 Disaggregate elasticity effects

Variables	Multinomial Logit Model			Mixed Spatially Correlated Logit Model		
	Self-elasticity	Cross-elasticity w.r.t. an adjacent zone	Cross-elasticity w.r.t. a non-adjacent zone	Self-elasticity	Cross-elasticity w.r.t. an adjacent zone	Cross-elasticity w.r.t. a non-adjacent zone
Log of zonal area (in mile^2)	1.06019	-0.01512	-0.01512	0.98157	-0.07767	-0.01904
Commute time (100's of minutes)	-1.28851	0.01837	0.01837	-1.4116	0.10670	0.00235
Percentage zonal Hispanic population interacted with Hispanic dummy variable	0.24650	-0.00351	-0.00351	0.29503	-0.23345	-0.00572
Absolute difference between zonal median income and household income ($100,000)	-0.00217	0.00003	0.00003	-0.00188	0.00015	0.00004
Population density (in 10 persons/mile^2)	0.22775	-0.00325	-0.00325	0.10869	-0.00204	-0.00025
Percentage of zonal area occupied by multifamily housing	-0.04791	0.00068	0.00068	-0.17917	0.01752	0.00030
Shopping accessibility	0.53541	-0.00763	-0.00763	0.4908	-0.03884	-0.00952

The self-elasticities in both MNL and MSCL models clearly indicate the dominant role played by commute travel time in residential choice modeling. The other important determinants of residential choice include zonal area and accessibility to shopping. As indicated earlier, the effect of commute travel time and the percentage of zonal area occupied by multifamily households is estimated to be higher in the MSCL model relative to the MNL model.

The difference in empirical results between the MNL and MSCL models suggests the need to apply formal statistical tests to determine the structure that is most consistent with the data. The models may be compared using a nested likelihood ratio test (the log-likelihood values at convergence for the two models are provided in the last row of Table 4). The result of such a

test leads to the clear rejection of the MNL model; that is, the test provides strong evidence that there is spatial correlation in residence choice and variation in responsiveness across households due to unobserved factors (the likelihood ratio test value is 25 which is larger than the chi-squared statistic with 4 degrees of freedom at any reasonable level of significance).

4.7 Summary

A MSCL model has been developed in this chapter for the analysis of location-related decisions of individuals and households. The MSCL structure is constructed by imposing the mixing structures on the proposed SCL model, which is essentially a GEV model customized for accommodating pair-wise spatially-correlated choice situations. The chapter submits, and demonstrates, that while the MMNL class of models is very general in structure, there are substantial computational efficiency gains to be achieved by using a mixed GEV structure. This is because the number of error components that needs to be specified in the MMNL structure to generate the desired spatial correlation pattern is very high for realistic location choice decisions. This leads to a high dimensionality of integration in the MMNL structure. In the empirical application described earlier, the use of a MMNL structure would entail a multidimensional integral of the order of 500, while the proposed MSCL model requires evaluation of only a three-dimensional integral.

In addition to computational efficiency gains, there is another more basic reason to prefer the MSCL model over an MMNL structure. This is related to the fact that closed-form analytic structures should be used whenever feasible, because they are always more accurate than the simulation evaluation of analytically intractable structures (see Train, 2002; pg. 191). In this regard, superimposing a mixing structure to accommodate random coefficients, over a closed form analytic structure that accommodates a particular desired inter-alternative error correlation

structure, represents a powerful approach to capture random taste variations and complex substitution patterns.

The empirical analysis in the chapter applies the MSCL model to examine the residential choice behavior of households in Dallas County using the 1996 Dallas-Fort Worth metropolitan area household activity survey. The empirical results indicate the important and dominant effect of commute travel time on residential location choice. Other variables significantly impacting residential choice include zone size, population density, percentage of zonal area occupied by multifamily housing, disparity between household income and median zonal income, percentage of Hispanic population for Hispanic households, and work and shopping accessibility. A comparison of the best specifications for the MNL and MSCL models indicates the significant presence of spatial correlation between contiguous zonal alternatives as well as differential responsiveness to exogenous variables across households. The MSCL model also leads to a statistically superior data fit. In addition, the results indicate that failing to accommodate spatial correlation and unobserved response heterogeneity can lead to incorrect conclusions regarding the elasticity effects of exogenous variables.

CHAPTER 5

ADDRESSING THE CONCEPT OF NEIGHBORHOOD

Because of the lack of micro-data, many past studies (the empirical analysis presented in Chapter 4 included) adopt the spatially aggregate approach of representing alternative locations by zones, as opposed to individual dwelling units, and measuring the locational attributes based on these zones. . However, the aggregate approach has a number of shortcomings. First, the use of zones as choice alternatives implies that only the choice of neighborhoods is considered. Unless the zones are internally homogenous, differences among the individual dwelling units and properties within the same zone are disregarded. Second, by examining all the spatial attributes over a single definitional configuration of zones, one assumes that every factor operates at one and the same spatial scale. Given the discussion presented in Section 3.2.2 about the neighborhood definition, this assumption is considerably unrealistic. Third, the model parameters are typically interpreted as the effects of the locational attributes on neighborhood choice. Yet, due to the presence of the MAUP, as discussed in Section 3.1.1, unless the zones are coterminous to the neighborhoods as perceived by residents, model estimates derived from arbitrarily defined zones do not correctly reflect the residents' choice behavior.

To overcome the abovementioned shortcomings of the aggregate approach, this chapter develops in Section 5.1 a disaggregate model, referred as the multi-scale logit (MSL) model, in which the choice alternatives are the individual housing units and each alternative is described by attributes measured at different spatial scales. Section 5.2 describes the application of the MSL model to households residing in the San Francisco Bay area. The empirical application demonstrates that, with the increasing availability of micro-level spatial data and the

technological advances in geographic information systems (GIS), the proposed disaggregate approach is not only possible, but also represents a more accurate and behaviorally realistic modeling approach than the grouped alternatives approach. The chapter concludes with some concluding remarks in Section 5.3.

5.1 Multi-Scale Logit Model

Given that the debate regarding the appropriate size and shape of spatially defined neighborhoods will not be resolved easily and, most likely, that no single unit of neighborhood will simultaneously satisfy the needs for measuring multiple neighborhood processes, one possible solution is to use multiple definitions of neighborhood within the same study. This solution has been implemented in, for example, hierarchical linear models for studying community psychology (Brodsky et al 1999, Ross et al 2000, Duncan et al 2003), housing price (Orford 2002) and, to a limited extent, urban form effect on travel behavior (Boarnet and Sarmiento 1998). To the best of the author's knowledge, the study by Quigley (1985), who incorporated in his model variables of commute times and racial composition measured at both the census tracts level and the town level, is the only residential location choice analysis to date that explicitly examined the effect of spatial factors at more than one scale.

The MSL model proposed in this study is a general structure for accommodating spatial attributes measured over a hierarchy of spatial definitions. The model considers each available dwelling unit as a choice alternative. The geographic location of an alternative i, as perceived by a household n, is described by a hierarchy of spatial units $S_{n,i}$. Let $X_s^{n,i}$ denote the vector of location attributes observed over a spatial unit s ($s \in S_{n,i}$) for household n of alternative i. The utility experienced by household n from choosing dwelling unit i is formally defined as:

$$U^{n,i} = \sum_{s \in S_{n,i}} \beta_s' X_s^{n,i} + \sum_{s \in S_{n,i}} \varepsilon_s^{n,i} . \quad (5.1)$$

In the above equation, the stochastic term, $\varepsilon_s^{n,i}$, represents any effects unobserved over a spatial unit s. Based on the simplifying assumption that the stochastic terms between different spatial scales are independent of each other, the stochastic terms with respect to a given household for a given dwelling unit are collapsed into a single term and the above equation becomes:

$$U^{n,i} = \sum_{s \in S_{n,i}} \beta_s' X_s^{n,i} + \varepsilon^{n,i} . \quad (5.2)$$

Furthermore, by assuming that each stochastic term $\varepsilon_{n,j}$ is IID Gumbel distributed, the unknown parameters, β_s, can be estimated using a MNL structure.

Compared to the grouped alternatives model, the MSL model structure provides a more realistic representation of how neighborhood is perceived as a hierarchy of ecological structures. The spatial hierarchy, $S_{n,i}$, can be configured differently to represent different hypothetical delineation of neighborhoods. For example, $S_{n,i}$ can be defined based on predefined administrative boundaries, such as the census geography, so that $S_{n,i} = S_i = \{\text{block}_i, \text{block-group}_i, \text{tract}_i\}$, where $\text{block}_i, \text{block-group}_i, \text{and tract}_i$ are the census block, block-group, and tract that contain alternative i, respectively. The spatial definition is thus objectively defined for all households. Alternatively, the concept of sliding neighborhoods and the fact that perceived boundaries are sometimes subjective motivate the use of a more ideal delineation consisting of circular units of varying radii centered about each alternative housing unit to describe neighborhood characteristics. Or, one can mix and match spatial units that reflect objectively and subjectively-defined boundaries. For instance, $S_{n,i}$ may comprise of the lot of the land on which dwelling i sits; the street block to which the dwelling belongs; the catchment area of the school where the children in household n would attend if they so choose to reside at

this dwelling; the half-mile or one-mile radius area around the dwelling where the socioeconomic and demographic composition would matter to the household; and the city and county to which the household would pay tax and from which the household would receive public amenity. As $S_{n,i}$ is indexed by n as well as i, the boundary definition may vary depending on households' characteristics.

Another strength of the MSL structure is that, by including neighborhood measures at more than one scale, the scale (or scales) at which each neighborhood factor operates is determined endogenously. That is, the model estimation process reveals not only the neighborhood determinants significantly influence the choice behavior, but also the spatial extent of their influence. By interpreting the parameters with reference to the spatial scale at which they are statistically significant, analysts gain insights about the spatial strengths, or cluster sizes, of various neighborhood processes.

5.2 Empirical Application

In order to demonstrate the advantages of the MSL model over the traditional single-level, grouped-alternatives approach, a number of models have been estimated using data from the San Francisco Bay Area, which consists of nine counties as shown in Figure 5.1. The first group of models estimated is of the single-level, grouped alternatives structure. One model is estimated for each of the census-defined geography: blocks, block groups, and tracts. The second groups of models are of the MSL structure. Two spatial hierarchical definitions are evaluated, one based on the hierarchy defined by census blocks, block groups and tracts (hereafter referred as the census-unit definition) and the other based on concentric circular areas of varying radii around each dwelling (hereafter referred as the circular-unit definition). The three single-level models are compared against each other to demonstrate the effects of the MAUP that renders the

model structure undesirable. The MSL models of the census-unit and of the circular-unit definitions are further compared against each other to examine (1) if and how the two configurations suggest different neighborhood effects on residential location choice behavior; and (2) which of the two (fixed versus sliding) is the 'better' neighborhood representation. The single-level specifications are compared against the census-unit MSL specification to highlight the advantages of the MSL structure over the conventional approach.

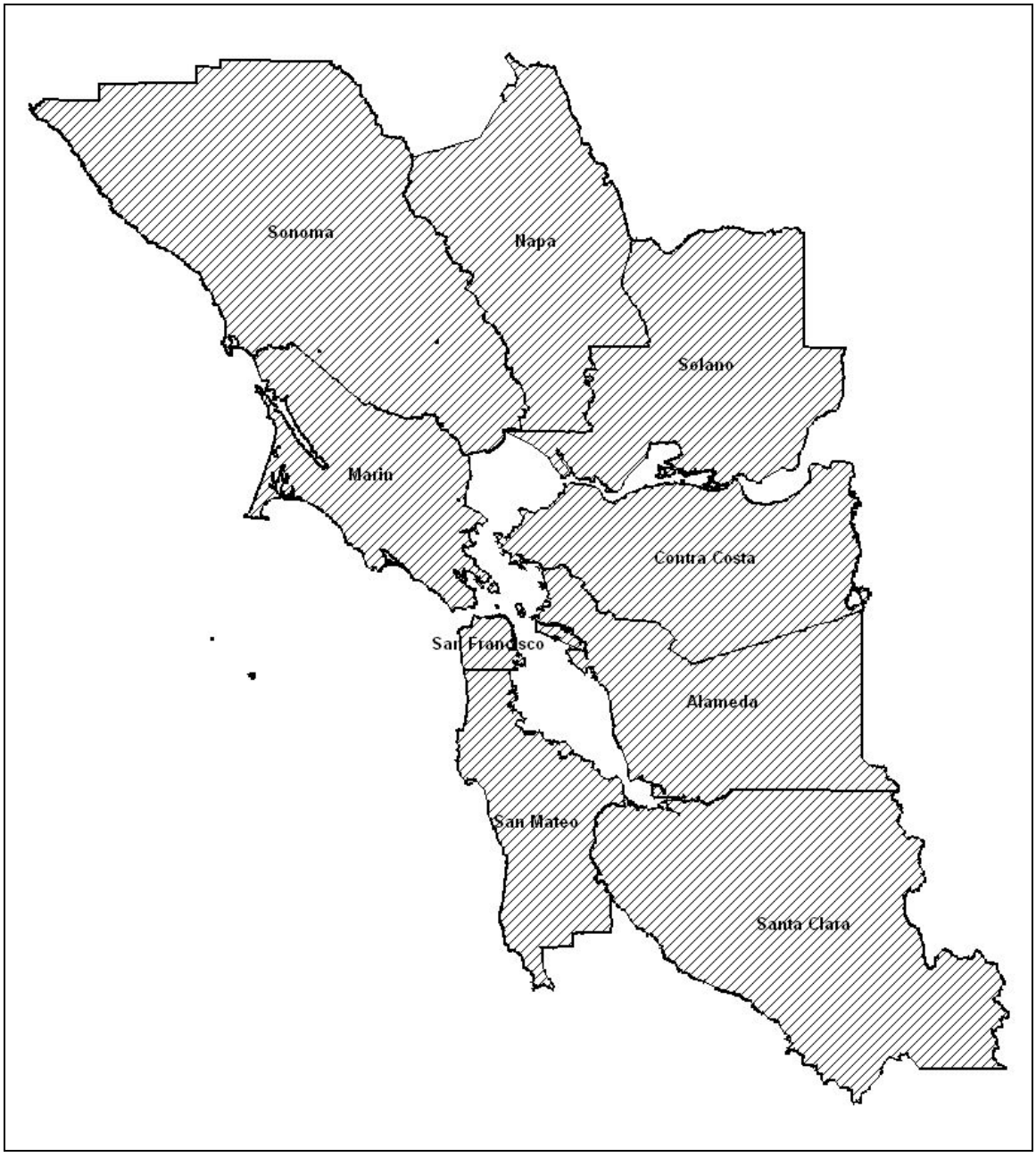


Figure 5.1 The study region covers the nine counties in the San Francisco Bay Area.

5.2.1 Data Source and Sample

The primary data source for the empirical application is the 2000 Bay Area Travel Survey (BATS) that collected, from members of 15,064 households, detailed information on individual

and household socio-demographic information, employment-related characteristics and all activity and travel episodes for a two-day period. The dataset also contains detailed geographical information, including the point geo-codes (in terms of longitude and latitude) of household residence from which the census block, block group and tract in which the residence situates can be identified. The geo-codes also make it possible to construct concentric circles of ½, 1, 1½ and 2 mile radii around each residence to form the circular-unit neighborhood definitions. From the surveyed households, 50% of those households living in single-family detached houses[2] are randomly selected to form the sample for model estimation. The sub-sampling eliminates the need to model neighborhood choice on the one hand and dwelling type choice on the other. Of course, this means the model will have nothing to say about the behavior of households who choose to live in apartments, duplexes or other types of housing. It also means that the model will correctly predict the effects of policy measures only if the measures have little effect on choice of dwelling type. Despite the limitation, the model will provide insights into the neighborhood preferences of residents of single-family detached houses and demonstrate the proposed methodology that can be applied to analyze the behavior of households who live in other dwelling types.

Following from the MSL structure, the choice alternatives faced by each household are defined as the individual dwellings. In theory, the universal choice set in this case comprises all the single-family detached houses in the Bay area. However, not only data about all such housing units in the area are unavailable, but it would be computationally impractical to consider them all. Therefore, the 4791 residences observed in the sample are assumed to be a random subset of such housing opportunities, and are representative of the unobserved choice set faced by each individual household. Based on the IID structure assumed about the stochastic terms, the model

[2] Single-family detached housing is the dominant type of housing among recent home-buyers in the US (National Association of Home Builders 2004).

can be consistently estimated by sampling of alternatives. The individual choice set thus constructed for each household includes the chosen alternative and nine randomly selected non-chosen alternatives.

In addition to the 2000 BATS data, a number of other data sources are used to derive measures about the choice alternatives. As listed in Table 5.1, these data sets provide information for different spatial scales and units. The Bay Area Metropolitan Transportation Commission provides employment and land-use distribution data for the TAZ in the Bay area. It also has information about the zone-to-zone distance and travel time by different mode of travel. The Census 2000 SF1, on the other hand, provides some demographic variables at the block level and some socio-demographic variables at the block-group level.

Table 5.1 Spatial variables considered in the residential choice models

Data source	Spatial level at which data is available	Variables considered
Bay Area Metropolitan Transportation Commission	Transport analysis zone	• Number of employment by sector (retail, wholesale, service, manufacturing, agriculture, and other) • Land-use acreage by purpose (residential, office, retail, and vacant)
Bay Area Metropolitan Transportation Commission	Transport analysis zone	Inter-zonal • Distances • Peak and off-peak travel times and costs by travel mode (car, shared ride, transit mode by both walk access, and transit mode by drive access)
Census 2000 population and housing data summary file 1 (SF1)	Census block	• Number of households • Population • Land/water area • Number of people by ethnicity (non-Hispanic Caucasian, African American, Asian, Hispanic, and other)
	Census block-group	• Median household income • Average household size • Number of housing units by size (single-family detached, apartments, etc) • Median housing value • Number of households by income quartiles

5.2.2 Data processing with a geographic information system

As data were not readily available for every level of the census- and the circular-unit hierarchies, a significant amount of effort has been devoted to process and assemble data to the desired format. The spatial data processing involved using a GIS software named TransCAD to perform the three steps described below.

Step (1): Because the most disaggregate level at which a subset of census variables are released are the block level, and the others at the block-group level, these variables need to be aggregated to the block-group level and the tract level to allow the estimation of single-level models for each levels of the census hierarchy. (See Figure 5.2)

Step (2): Employment and land-use data are available only for the TAZ, and not the census units. Therefore, in order for these attributes to be considered in the single-level census-based models, these TAZ level data must be overlaid to the census units. The overlay operation requires assumptions to be made about the distribution of the TAZ attributes. For this analysis, the TAZ attributes are assumed to follow the uniform distribution within each zone so that data for a given zone can be disaggregated uniformly over the zone. For instance, if the number of service employments in a 10 squared-mile zone is 100, then every squared-mile area in the zone is assumed to have 10 service employments. The disaggregated data are then projected onto, and re-aggregated over, each of the three census layers to produce the corresponding measures for the census blocks, block groups and tracts. (See Figure 5.3)

Step (3): The overlay operation described in Step (2) is repeated to generate measures for the circular units defined for each dwelling alternatives. In this case, the source of overlay includes the TAZ-level data and the census data available for the blocks and block-groups. The target layer for overlay is the layer containing the circular units. (See Figure 5.4)

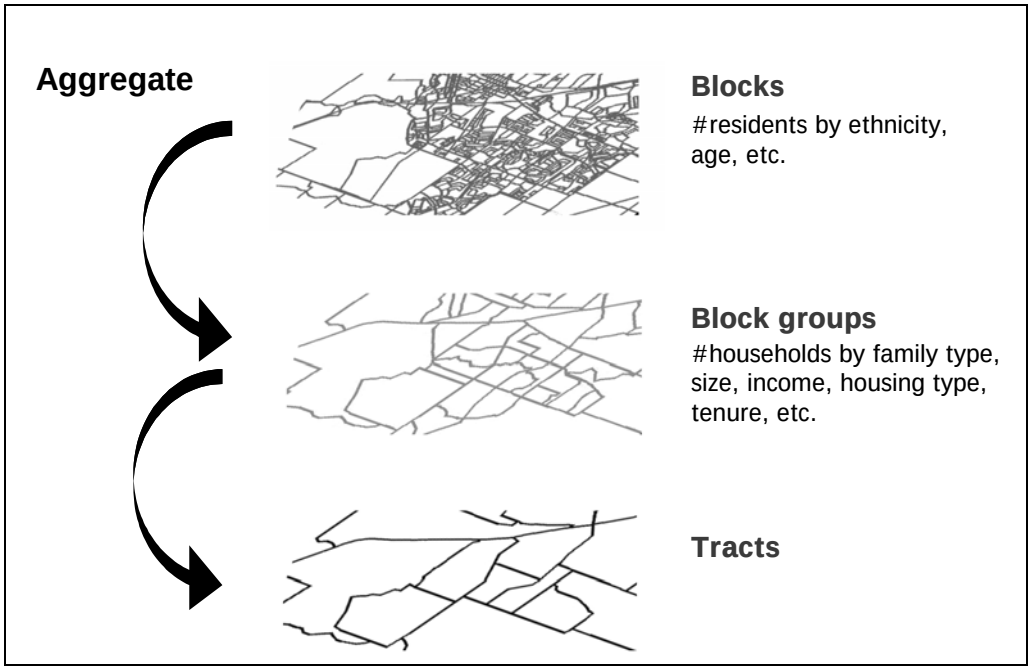


Figure 5.2 Spatial data processing: step (1) - aggregating data down the census hierarchy

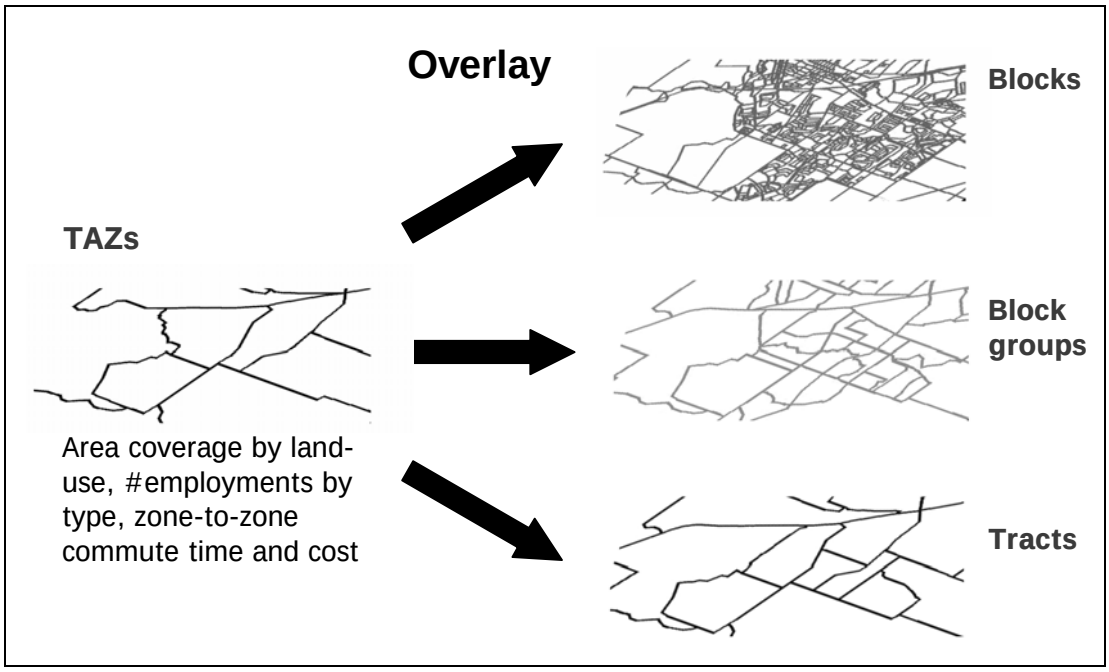


Figure 5.3 Spatial data processing: step (2) - Overlaying TAZ data onto the census units

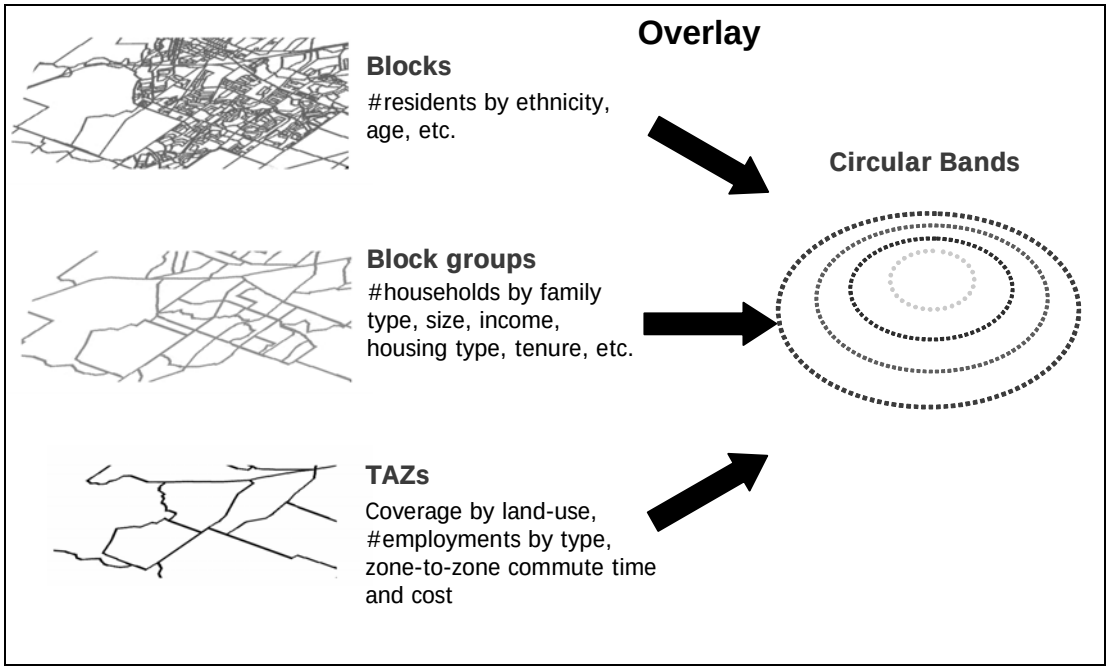


Figure 5.4 Spatial data processing: step (3) - Overlaying census and TAZ data onto the circular units

5.2.3 Variable Specifications

After data are obtained for the desired spatial definitions, they are used to compute the following sets of variables.

5.2.3.1 Commute-related variables

Based on the residents' work and alternative residential TAZ locations, the auto commute time and cost variables are extracted from the level-of-service data. These variables are then interacted with individual work status, gender, ethnicity and income variables. Due to the high correlation between the time and cost variables, commute cost variables were excluded from the final specification (between the commute time and commute cost variables, the former performed much better).

5.2.3.2 Regional accessibility variables

The regional accessibility measures for shopping, recreational, and employment activities are computed as follows:

$$A_i^{Shop} = \sum_{j=1}^{N} \frac{1}{N} \frac{R_j}{d_{ij}},\ A_i^{Emp} = \sum_{j=1}^{N} \frac{1}{N} \frac{E_j}{d_{ij}}, \text{ and } A_i^{Rec} = \sum_{j=1}^{N} \frac{1}{N} \frac{V_j}{d_{ij}}$$

where A_i^{Emp}, A_i^{Rec} and A_i^{Shop} denote the shopping, employment and recreational accessibility indices, respectively, for TAZ i; R_j, E_j and V_j are the number of retail employment, number of basic employment and vacant land acreage in TAZ i, respectively; d_{ij} is the distance between zones i and j. Due to data constraints, these zonal accessibility measures are used in the subsequent analysis as proxy for point-to-region accessibility measures for each observed residence. The accessibility measures are also interacted with household income, structure and ethnicity to test for households' differential sensitivity. A few points to note here. Large values of the accessibility measures indicate more opportunities for activities in close proximity of that residence, while small values indicate residences that are spatially isolated from such opportunities. In addition, in measuring recreation accessibility, the vacant land acreage is used in lieu of parkland acreage.

5.2.3.3 Socioeconomic and demographic variables

Several variables are computed to test for the presence of segregation along various socioeconomic and demographic dimensions. These include the racial composition variables (percentage of population by race), household type composition variables (percentage of households by type), tenure composition variables (percentage of households owning or renting), household income homogeneity (absolute difference between household income and zonal median income), and household size homogeneity (absolute difference between household size

and zonal average household size). The racial, household type and tenure composition variables are further interacted with the respective household attributes.

5.2.3.4 Neighborhood design variables

A variety of neighborhood design (land-use) measures were considered for this analysis. These include density measures (population density, housing density), land-use composition measures (percentage of coverage by land-use type) and employment density measures (number of employment per person). Also considered is a more complex measure of land-use diversity defined by:

$$LUMIX_s = 1 - \frac{\left|\frac{R_s}{T_s} - \frac{1}{4}\right| + \left|\frac{C_s}{T_s} - \frac{1}{4}\right| + \left|\frac{I_s}{T_s} - \frac{1}{4}\right| + \left|\frac{O_s}{T_s} - \frac{1}{4}\right|}{\frac{3}{2}}$$

where T_s is the total area of the unit of analysis s; and R_s, C_s, I_s and O_s are the acreage of residential, commercial, industrial and other land use type. This land-use mix index takes a value between 0 and 1, where 1 indicates perfect mixing of land uses and 0 indicates that the land in a particular area is completely dedicated to a single land use (see Bhat and Gossen, 2002). Despite the interest and claims of the advocates of mixed land-use communities (a feature of the New Urbanism design), there is little information in the literate about the effects of mixed land-uses on residential location choice, or about the suitable balance between different types of land-uses within neighborhoods. Hence, unlike most of the variables discussed in Section 2.6.4, the expected sign for some of these neighborhood design variables is not intuitively known.

5.2.3.5 Other variables

In addition to the four groups of variables described above, a set of geographic indicators are also considered in the estimation process. These indicators take the form of county-specific

dummy variables to capture the average effects of any unobserved attributes at the county level. However, no such effects were found to be statistically significant. Also, crime rates (measured at the county level) were also considered, but excluded from the final specifications due to lack of statistically significant impacts. This does not, however, necessarily reject the hypothesis that safety is not an influencing factor on residential utility. Rather, it is likely that the counties are too broad a spatial scale to reflect safety considerations. Crime rate statistics at a finer spatial resolution would be very helpful, but were not available.

5.2.4 Estimation Results

5.2.4.1 Results for single-level models

Three MNL models are estimated, each using a single level of the census geography. It should be noted that, since the commute-related and the accessibility-related measures derived from the TAZ data are used to approximate the point-to-region commute and accessibility, the same measures are used for the estimation of all three models. Only the measures relating to segregation effects and neighborhood designs are compiled differently for the different census scales. The parameter estimates and t-statistics are presented in Table 5.2. The similarities and differences among the model estimates are discussed below.

Table 5.2 Estimation results for the single-level Models

Variables			Block Level Model		Block-group Level Model		Tract-Level Model	
Location attribute	(x	Household attribute)	Param.	t-stat.	Param.	t-stat.	Param.	t-stat.
Commute-related variables								
Commute Time	x	Full-time male workers	-0.0432	-9.87	-0.0421	-9.43	-0.0425	-9.49
Commute Time	x	Full-time female workers	-0.0570	-12.30	-0.0552	-11.74	-0.0561	-11.87
Commute Time	x	Part-time male workers	-0.0752	-9.00	-0.0731	-8.62	-0.0731	-8.58
Commute Time	x	Part-time female workers	-0.0912	-16.23	-0.0868	-15.16	-0.0874	-15.19
Commute Time	x	Caucasian household	-0.0095	-3.10	-0.0103	-3.33	-0.0104	-3.35
Commute Time	x	household income (in $100,000)	0.0067	2.06	0.0060	1.80	0.0063	1.91
Accessibility-related variables								
Employment Accessibility			-0.0154	-9.43	-0.0142	-6.49	-0.0163	-6.94
Employment Accessibility	x	single-person household			0.0096	5.29	0.0094	4.92
Employment Accessibility	x	Couple only household	0.0016	9.48			0.0052	2.43
Employment Accessibility	x	household income (in $1,000)	0.0046	1.82	0.0012	6.11	0.0017	5.61
Shopping Accessibility	x	Couple only household	-0.0192	-3.08	0.0139	2.08		
Shopping Accessibility	x	household income (in $1,000)					-0.0019	-2.30
Residential segregation effects								
Share of Caucasian population	x	Caucasian household	0.6504	7.00	0.8231	6.61	0.7710	5.95
Share of African population	x	African household	4.6721	11.77	5.1447	11.45	5.1719	11.09
Share of Asian population	x	Asian household	9.4455	9.33	10.5500	7.48	10.1809	6.27
Share of Hispanic population	x	Hispanic household	4.0716	6.24	4.6325	5.52	5.2853	5.94
Share of other ethnic population	x	African household			23.9682	2.44	34.0245	2.80

Share of other ethnic population	x	Hispanic household	6.1827	1.84	20.6790	2.65	34.1936	3.33
Share of couple-only household	x	couple-only household			1.5846	4.92	1.8218	3.69
Share of nuclear-family household	x	nuclear-family household			1.1263	3.29	0.8170	1.69
Absolute difference between zonal median income and household income (in $1,000)					-0.0099	-11.74	-0.0089	-9.43
Absolute difference between zonal average household size and household size					-0.1877	-4.60	-0.1855	-3.58
Share of owner-occupied housing	x	own house					0.2437	1.69
Zonal Median housing value (in $1,000)	x	Inverse of household income			-0.0217	-3.90	-0.0215	-3.68
Neighbourhood design factors								
Density over land area (per 10,000 mi^2)			0.0857	2.77	0.2482	4.17	0.3231	4.77
Density over land area (per 10,000 mi^2)	x	household income > $75,000	-0.1237	-2.84	-0.1502	-2.15	-0.1762	-2.37
Density over land area (per 10,000 mi^2)	x	African household	0.2677	3.03				
Density over land area (per 10,000 mi^2)	x	Couple only household	-0.0989	-2.15	-0.1725	-2.02	-0.1678	-1.82
Density over land area (per 10,000 mi^2)	x	nuclear-family household			-0.2444	-2.87	-0.2423	-2.58
Share of commercial land-use			2.5308	7.89	1.2615	3.65	1.5260	4.20
Share of commercial land-use	x	Couple only household	-0.8941	-2.82	-0.8056	-2.30	-0.7356	-1.97
Share of commercial land-use	x	household income (in $1,000)	-0.0267	-7.83	-0.0122	-3.31	-0.0142	-3.68
Share of residential land-use	x	single-person household	0.6138	4.61				
Share of residential land-use	x	household income (in $1,000)	-0.0018	-2.51	-0.0016	-2.19		
Number of retail employment (per 10 people)	x	Single-person household	0.0035	2.67				

Number of retail employment (per 10 people)	x	Couple only household			0.0009	2.50		
Number of service employment (per 10 people)	x	Nuclear family household					0.0244	1.66

Commute-related variables

As shown in Table 5.2, the best specification found for all three models includes the commute-time measures interacted with gender, employment level (full time versus part time), race, and income. The parameter estimates indicate that households tend to locate themselves closer to, rather than farther from, the work locations of the workers in the household. In particular, households locate themselves close to the workplace of the female workers in the household. This gender disparity is independent of employment level. A similar higher responsibility hypothesis may be the underlying cause for the greater commute time effect of part-time workers relative to full-time workers. The racial disparity in commute sensitivity indicates greater spatial job-housing mismatch for non-Caucasians compared to Caucasians. The positive sign associated with the interaction of commute time and income may be a reflection of the willingness of higher income earners to travel further in exchange for better housing quality. The magnitudes and signs of the commute-related parameter estimates are very consistent across the three models.

Accessibility-related variables

The coefficients on employment accessibility for all three models have a negative sign, suggesting households' general aversion to locations too close to substantial employment centers after the direct accessibility to work locations is accounted for. However, taken together with the parameter on the interaction term between employment accessibility and income, the results indicate that households have earnings higher than a certain threshold tend to locate themselves near employment centers. The three models show different preference patterns across household types relating to regional employment accessibility. The preference for good access to employment opportunities is confined to couple-only households in the block-level model, but to single-person households in the block-group level model. The tract level model indicates that,

while single-person and couple-only households both prefer higher employment accessibility, the propensity is higher for single-person households than couple-only households.

Different effect of regional shopping accessibility are suggested by the three models. While the block-level model suggests that, compared to other household types, couple-only households are less likely to locate in areas with good access to shopping opportunities, the block-group level model suggests the contrary. Furthermore, only the tract-level model shows significant interactive effect of shopping accessibility with household income level. Higher income households are less likely to locate in areas close to retail centers. Recreation accessibility measures are absent from all three specifications, suggesting that proximity to recreational opportunity does not influence residential choice behavior.

Residential segregation effects

A consistent finding among all three models is the evidence of substantial racial segregation. The models also find that Hispanics and/or African Americans are more likely to reside near other minority groups.

The block-group level and the tract level models are further comparable in terms of additional segregation-related variables that were unavailable for the block level. The income dissimilarity parameter confirms the income segregation hypothesis found in literature. In addition to segregation by race and socioeconomic status, households tend to cluster with other households of similar structure as suggested by the parameters associated with household size dissimilarity and the household type composition. More specifically, couple households tend to locate themselves in neighborhoods (block groups, tracts) populated by other couple households, while the clustering of nuclear-family households is also evident at both levels. The sign associated with the price-income ratio has the expected negative sign and suggests that affordability is less of an issue for high-income households.

Neighborhood design factors

The consistency of the density-related parameters is mixed among the three models. Density, density interacted with income, and density interacted with couple-only households are found to be significant in all specifications. The block-level model is the only one showing that African-Americans are the racial group most likely to reside in high density areas. The block-level model, however, is also the only model that does not show nuclear-family households' aversion to high population density.

Significant effect of commercial land-use on residential choice is found in all three models. Taken together, the parameters suggest that the attractiveness of local access (within the same census block, block-group, or tract) to stores diminishes as household income increases. In addition, couple-only households are less drawn to areas with high commercial land-use than other types of households. Surprisingly, the preference for residential-oriented areas is found only for the single-person households in the block-level model. The sign associated with percentage of residential land-use interacted with income suggests that, in general, households are less likely to reside in areas of high percentage of residential land-use, with the aversion diminishes with increased income. This is perhaps attributed to the preference of high-income households for suburban, residential-oriented areas.

The three models suggest very different effects of employment densities. The block level model indicates that only the single-person households are more likely to locate in blocks with high retail employment density. The block-group level model suggests that couple-only households are more likely to locate in block-groups with high retail employment density. The tract level model suggests that only the nuclear-family households are drawn to good local access (within the census tract) to services.

It is surprising, and perhaps counter-intuitive, that the land-use mix parameter, either by itself or through interaction with the various household characteristics, does not appear to be a significant factor in all three models. This suggests that, after the access to particular types of land-use (commercial) or amenity (service) is accounted for, households are not particularly attracted to mixed land-use.

5.2.4.2 Results for the MSL models

This section describes the estimation results for a census-unit definition model and a circular-unit definition model of the MSL structure. The results of the final specification are presented in Table 5.3 (census-unit definition) and Table 5.4 (circular-unit definition) and discussed below by variable group. Overall the census-unit and the circular-unit models are consistent in the signs of the parameter estimates. The final specifications differ in the presence and the absence of certain variables and the spatial level at which these variables are significant.

Commute-related variables

The two models are most comparable in terms of the parameter estimates (both in sign and magnitude) relating to commute time (see Table 5.3 and Table 5.4). The same gender and racial disparities in commute sensitivity revealed by the single-level models are found in both MSL models.

Table 5.3 Estimation results for the census-unit MSL model

Variables			Multi-scale					
Location attribute	(x	Household attribute)	Param.			t-stat.		
Commute-related variables								
Commute Time	x	Full-time male workers	-0.043			-9.55		
Commute Time	x	Full-time female workers	-0.056			-11.88		
Commute Time	x	Part-time male workers	-0.073			-8.57		
Commute Time	x	Part-time female workers	-0.088			-15.26		
Commute Time	x	Caucasian household	-0.010			-3.21		
Commute Time	x	household income (in $100,000)	0.006			1.89		
Accessibility-related variables								
Employment Accessibility			-0.016			-6.76		
Employment Accessibility	x	Household income (in $1,000)	0.001			6.10		
Employment Accessibility	x	Single-person household	0.009			4.25		
Shopping Accessibility	x	Couple only household	0.003			1.91		
			Block		**Block group**		**Tract**	
			Param.	t-stat.	Param.	t-stat.	Param.	t-stat.
Residential segregation effects								
Share of Caucasian population	x	Caucasian household	0.667	6.72				
Share of African population	x	African household	9.426	8.97				
Share of Asian population	x	Asian household	2.887	3.92	2.348	2.87		
Share of Hispanic population	x	Hispanic household	1.737	1.66	3.467	2.69		
Share of other ethnic population	x	African household					38.606	2.98
Share of other ethnic population	x	Hispanic household					32.074	3.14
Share of single-person household	x	Single-person household					0.962	2.10
Share of couple-only household	x	Couple-only household			0.987	2.39	1.286	2.14
Share of nuclear-family household	x	Nuclear-family household			1.136	3.23		
Absolute diff. between zonal median income and household income ($1,000)					-0.010	-11.45		
Absolute diff. between zonal average household size and household size					-0.185	-4.23		

Share of owner-occupied housing	x	Own house					0.234	1.67
Zonal Median house value ($1,000)	x	Inverse of total income			-0.020	-3.56		
Neighbourhood design factors								
Density (per 10,000 mi^2)							0.227	3.70
Density (per 10,000 mi^2)	x	African American household	0.182	2.00				
Density (per 10,000 mi^2)	x	Nuclear-family household			-0.161	-2.06		
Density (per 10,000 mi^2)	x	Household income > $75,000					-0.240	-3.28
Share of commercial land-use					1.283	3.59		
Share of commercial land-use	x	Couple only household			-0.712	-1.92		
Share of commercial land-use	x	Household income (in $1,000)			-0.012	-3.12		
Share of residential land-use	x	Single-person household	1.202	2.08	-1.118	-1.88		
No. of service employment (per 10 people)							-0.055	-2.30
No. of service employment (per 10 people)	x	Single-person household					0.058	2.40
No. of service employment (per 10 people)	x	Couple only household					0.059	2.28
No. of service employment (per 10 people)	x	Nuclear family household					0.080	2.90
Number of observations			4791					
Mean log-likelihood at convergence			-0.187306					

Table 5.4 Estimation results for the circular-unit MSL model

Variables			Multi-scale					
Location attribute	(x	Household attribute)	Param.			t-stat.		
Commute-related variables								
Commute Time	x	Full-time male workers	-0.042			-9.25		
Commute Time	x	Full-time female workers	-0.055			-11.74		
Commute Time	x	Part-time male workers	-0.072			-8.51		
Commute Time	x	Part-time female workers	-0.087			-15.07		
Commute Time	x	Caucasian household	-0.011			-3.37		
Commute Time	x	household income ($100,000)	0.006			1.65		
Accessibility-related variables								
Employment Accessibility			-0.012			-5.45		
Employment	x	Household income ($1,000)	0.001			5.07		
Employment	x	Single-person household	0.009			4.90		
			R = 0.5 mi		**R = 1.0 mi**		**R = 1.5 mi**		**R = 2.0 mi**	
			Param	t-stat.	Param	t-stat.	Param	t-stat.	Param	t-stat.
Residential segregation										
Share of Caucasian	x	Caucasian household	0.778	6.11						
Share of African	x	African household	10.338	7.32						
Share of Asian	x	Asian household	5.184	11.25						
Share of Hispanic	x	Hispanic household	5.294	5.72						
Share of other ethnic	x	African household			47.128	3.79				
Share of other ethnic	x	Hispanic household	32.212	3.35						
Share of single-person	x	Single-person household	1.206	2.54						
Share of couple-only	x	Couple-only household	2.214	5.41						
Share of nuclear-family	x	Nuclear-family household	1.023	2.22						
Absolute difference between zonal median income and household income (in $1,000)			-0.012	-11.73						

Absolute difference between zonal average household size and household size				-0.213	-3.92				
Share of owner	x	Own house						0.631	3.31
Zonal Median housing value ($1,000)	x	Inverse of household income	-0.023	-3.71					
Neighbourhood design factors									
Density (per 10,000 mi^2)			0.153	2.75					
Density (per 10,000	x	Nuclear-family household			-0.216	-1.97			
Share of commercial land-use					1.388	2.93			
Share of commercial land	x	Household income ($1,000)			-0.012	-2.50			
Number of service employment (per 10 people)	x	Single-person household					0.113	2.17	
Land-use mix	x	Own no cars					-7.513	-2.21	9.178	2.64
Land-use mix	x	Own 1 car							0.506	1.78
Land-use mix	x	Has children							-0.575	-1.98
Land-use mix	x	Has seniors			-0.660	-1.88			0.666	1.84
Number of observations			4791							
Mean log-likelihood at convergence			-0.18773							

Accessibility-related variables

The two models also have similar estimates for the sensitivity to employment accessibility. Taken together with the parameter on the interaction term between employment accessibility and income, the results indicate that households earning an annual income greater than $16,000 (in the census-unit model) or $12,000 (in the circular-unit model) tend to locate themselves near employment centers even after the direct accessibility to work locations has been accounted for. Single-person households are also found to prefer closer proximity to regional employment opportunities than other types of households. The effect of regional shopping accessibility is different for the two models. While the census-unit model suggests that, compared to other household types, couple-only households prefer good access to shopping opportunities, the circular-unit model suggests it is the single-person households who have such preference. Similar to that found for the single-level models, recreation accessibility does not appear to be an influencing factor for residential location choice.

Residential segregation effects

A consistent finding in both models is the evidence of residential segregation across a number of dimensions. As indicated by the magnitude of the parameter estimates, African-Americans are the most segregated group, while Hispanics and Asians are segregated to a lesser degree – this finding is in agreement with the national demographic trend found in Iceland et al (2002). Compared to Caucasians and Asians, Hispanics and African-Americans are much more likely to integrate with other minority groups. It is also interesting that, despite the common impression of strong Black-White segregation, the coefficient associated with the share of Caucasian population interacted with Caucasian household is relatively much smaller. This perhaps indicates that the Caucasians in the Bay area have a relatively high tolerance for the presence of the other population groups as a whole in their neighborhoods.

Evidence of strong racial segregation is a common finding in past studies of residential location choice. What past models of residential choice have not been able to reveal is the differential spatial extents of the racial clustering behavior. The MSL models show that the size of racial clusters does vary for different racial groups and that different neighborhood definitions suggest different cluster sizes. In Table 5.4, almost all the racial segregation variables are significant at, and only at, the 0.5-mile radius level, with the exception being the aggregation of the 'other' ethnic population with African-American households. This suggests that racial clusters are generally of 0.5 mile in radius. The census-unit model in Table 5.3, however, tells a different story. The aggregation among Caucasians and among African-Americans is prominent in census blocks only; whereas the aggregation among Asians and among Hispanics is found in both census blocks and block-groups. Also, the integration of African-Americans and of Hispanics with the other minority groups is found only at the tract level. This difference in spatial scales of racial segregation between the two models is perhaps a result of the MAUP, the variation in the sizes of census units, or other factors that require further exploration and research to uncover.

Clustering of households of similar structure is suggested by the parameters associated with household type composition and household size homogeneity. On the one hand, Table 5.4 shows that single-only, couple-only and nuclear-family households tend to locate themselves in neighborhoods (within 0.5mi radius) populated by other single-only, couple-only and nuclear-family households, respectively. On the other hand, Table 5.3 suggests the presence of clustering among couple-only and nuclear-family households at the census block-group level, and clustering among single-only and couple-only households at the census tract level. Again, the observed inconsistency between the two models regarding the clustering of household structure calls for further research. In addition to segregation by race and household structure, households are

found to locate near other households of similar income level, confirming the income segregation hypothesis found in the literature.

Neighborhood design factors

Interestingly, the consistency of the neighborhood design parameters between the census-unit model and the circular-unit models is very mixed. Density and density interacted with nuclear-family households are two of the few variables that are significant in both specifications. Yet, the spatial extents of influence for the two variables are quite different between the two models. The census-unit model implies that households generally prefer census tracts of high population density but nuclear-families show an aversion to census block-groups of high density. The circular-unit model indicates that households generally have an affinity for high population density, but only within close (0.5mi) proximity of their residence. Population density has a negative influence on nuclear-family households and the extent of this influence is within a 1 mile radius of their residence. In the census-unit model, density is also found to have an additional positive effect on African-Americans and a negative effect on high-income households.

Of the several land-use composition variables and their interaction terms with household characteristics, only the share of the commercial land-use variable and the share of commercial land-use interacted with income are significant in both the census-unit and the circular-unit model specifications. Taken together, the parameters suggest that the attractiveness of local access (within the same census block-group, or within the 1 mile radius neighborhood) to stores diminishes as household income increases. In addition, as suggested by the census-unit model but not the circular-unit model, couple-only households are less drawn to areas (census block-groups) with high commercial land-use than other types of households. The census-unit model also indicates opposite effects of residential land-use at two spatial scales on single-only households.

As for the employment density variables for several employment sectors, only service employment density has an influence on residential choice decisions. The census-unit model indicates that, on the one hand, nuclear-family, couple-only and single-only households are more likely to locate in census tracts with high service employment density, with the degree of likelihood being the highest for nuclear-families and the lowest for single-only households; but on the other hand, households other than these three types show aversion to such tracts. The circular-unit model suggests that only the single-person households are drawn to good local access (within 1.5 mile radius) to services.

Similar to that observed for the single-level models, the land-use mix parameter and its interaction with the various household characteristics do not appear to be a significant factor in the census-unit model. With the circular-unit model, however, the effect of land-use mix on residential location choice is statistically significant. Measured within a 2 mile radius around the residences, heterogeneous land-use composition has a positive effect on households with zero or only one car, and households with senior citizens. It also has a negative effect on households with children. These effects appear to be intuitive. However, the land-use mix parameters also show negative effects on households with no cars when measured within a 1.5-mile radius, as well as on households with senior citizens when measured within a 1-mile radius. These results might not be intuitive and are perhaps due to the high correlation between the values of the land-use mix index measured at different scales.

Measures of fit

Since the number of variables present in the final specifications is different between the census-unit model and the circular-unit model, the log-likelihood ratios are not directly comparable. Instead, the goodness-of-fit of the two models are measured and compared using the adjusted likelihood ratio. The census-unit model is found to be marginally superior to the

circular-unit model. Despite being statistically inferior, however, the circular-unit model has its conceptual merits: (1) it represents the notion of 'sliding neighborhood', which, in the context of residential location choice, probably better reflects households' perception than 'fixed neighborhood'; (2) it gives a more tangible indication of 'how far' into space a neighborhood attribute matters, rather than 'what census unit' matters; (3) though both models show the general trend of demographic factors having smaller areas of influence than the neighborhood design factors, the circular-unit model is more consistent in showing so than the census-unit model.

5.2.4.3 Single-level versus multi-scale

As found in Table 5.2 and discussed in Section 5.2.4.1, the three grouped alternatives models estimated based on different levels of the census geography imply differing residential choice behaviors, raising the questions of which of the models is the 'true' model and based on which of these models spatial policies or residential development decisions should be made. The difference in the modeling results is at least partly attributed to the MAUP. It is likely that none of these models is truly reflect the residential choice behavior of the residents of single-family housings in the Bay area.

Statistically, based on the adjusted likelihood ratio test, the three single-level models are inferior than the MSL estimated using data measured for all three census scales. The MSL model structure is a more realistic representation of how a neighborhood is perceived as a hierarchy of residential groupings. The MSL structure is also advantageous over the single-level models in that it allows the spatial extent of influence of each variable to be determined endogenously. By interpreting the parameters with reference to the spatial scale at which they are statistically significant, interesting observations can be drawn about the clustering behavior underlying residential choice decision-making.

5.3 Summary

This chapter has described the proposed MSL model structure, in which spatial attributes are considered at multiple scales, as opposed to the conventional approach based on a single scale. Thus, the MSL structure more realistically represents, as compared to the conventional approach, how a neighborhood is perceived as a hierarchy of residential groupings. The hierarchy of spatial units used for measuring location attributes can reflect objectively defined neighborhood boundaries, subjectively defined neighborhood boundaries, or both. The boundary definitions can vary for different dwelling units, as well as for different households to reflect their difference in perception. Although the MSL model does not readily provide answers as to what is the 'correct' spatial delineation to use, it serves as a useful tool for exploring alternative hierarchical representations. The structure is also valuable in that it allows the variables' spatial extent of influence be determined endogenously. By interpreting the parameters with reference to the spatial scale at which they are statistically significant, insights are gained about the underlying clustering of observations.

The empirical analysis in the chapter demonstrated the shortcomings found in the grouped-alternatives modeling approach, particularly the effect of the MAUP. Based on data measured using census geography, the hierarchical nature of the MSL model is shown to outperform the single-level models both conceptually and statistically. The MSL model produced richer and more interpretable results than with the grouped-alternatives approach. The MSL structure is also applied to empirically test the implementation of 'fixed neighborhood' (the census units) against that of 'sliding neighborhood' (the concentric circular units). A number of conclusions can be made from the empirical results. First, the census-unit and the circular-unit models are generally consistent in the signs and magnitudes of the parameter estimates relating to the point-to-region measures, including the commute variables and the regional accessibility

variables. For the other variables considered in the analysis, the two models differ in the variables that are significant and the spatial scale at which these variables are significant. Second, for parameters that are found to be significant in both models, they tend to have the same sign but their respective values can differ up to 300%. Third, both models suggest that the social-economic and demographic composition variables operate mostly at a lower spatial level (up to the block-group level or up to 1 mile around each residence), while the influence of the land-use variables can be significant over an entire census tract or an area of 2 mile radius. The aforementioned findings can perhaps explain why previous residential choice studies utilizing the grouped alternatives model sometimes fail to provide empirical evidence for certain intuitive hypotheses about the impact of neighborhood characteristics on residential utility. These findings may also help improve our ability to disentangle the relationship between urban form and travel with regard to the self-selection problem (Boarnet and Sarmiento 1998, Crane 2000 and Lund 2003). This is because researchers inquiring into the urban form and travel connection are accustomed to using a single set of fixed, administratively defined boundaries. For example, Boarnet and Sarmiento (1998) attempt to account for the self-selection problem by introducing demographic and housing stock variables (as instruments) at a level of geography similar to that of the land-use variables. Given the findings from this paper, the fact that the spatial factors may operate at different scales probably contributes to their mixed observations about the validity of those instrumental variables.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Summary

An understanding of why, who, and where questions associated with residential choices is important to the researchers and practitioners who are inclined to seek land use solutions to transportation problems. The potential of any single policy, such as jobs-housing balance or New Urbanism design, to help meet the needs of diverse populations is limited by a wide range of factors influencing households' decisions about residential location. Nothing would be gained from implementing a spatial policy if its effect was diminished by households' other considerations and therefore could not result in the desired residential pattern. Accurate models of residential location choice are therefore valuable tools to help devise effective spatial policies.

Over the last two decades, there have been limited advances in the conceptualization of, and modeling methodology for, the residential location choice problem. The conventional approach is to assume that, behaving based on the RUM principle, decision makers select residential locations in the same manner as they would select discrete commodities such as the brand of coffee or the mode for travel. The approach typically considers administratively defined zones, which represent spatial groupings of individual dwellings, as the commodity for consumption.

This dissertation research is based on the hypothesis that there are unique features of the residential choice problem that distinguish it from other types of discrete choice problems. Any analysis that hopes to provide a robust explanation for residential choice dynamics and a framework for evaluating housing policy must take seriously these distinguishing features. Two

important spatial features of the residential choice problem are addressed in this study. The first feature relates to the perceived similarity between neighboring choice alternatives that are intangible or difficult to quantify. Failure to account for such inter-alternative correlations would result in biased parameter estimates for the conventional MNL models, leading to inaccurate interpretations of choice behavior. To address the problem, this dissertation proposed the MSCL structure that is capable of capturing unobserved spatial correlation between neighboring residential alternatives as well as unobserved heterogeneity across households. The MSCL represents a powerful approach that combines the state-of-the-art developments in closed-form GEV models with the state-of-the-art developments in open-form mixed distribution models.

The second spatial issue addressed in this dissertation is the representation and measurement of spatial factors. By measuring spatial factors over administratively defined zones, the conventional grouped alternatives approach fails to relate the configuration of spatial units to decision makers' perception of space, subjecting the modeling results to effects of the MAUP. The dissertation has proposed the MSL model structure for representing spatial factors over a hierarchy of spatial definitions, as opposed to the conventional 'flat' approach. The modeling approach is innovative in that it allows the choice factors' spatial extent of influence be determined endogenously. In addition, the hierarchy of spatial definitions can be configured differently to represent hypothetical delineation of neighborhoods as perceived by different households for different residential alternatives. The MSL model can therefore be used to explore alternative hierarchical spatial representations.

The proposed models have been estimated and their merits empirically demonstrated using revealed preference data*. A variety of hypotheses about the factors which might affect the

* Note that the use of revealed preference data means the results represent the preferences as well as the external constraints on location choice. Since the time the households moved into their current residence

residential decision were examined. These factors include commute-related factors, regional accessibility measures, factors contributing to residential segregations, neighborhood design and land-use distribution factors, school quality, and safety factors. The empirical comparison of the MSCL model against the conventional MNL model showed significant presence of spatial correlation between contiguous zonal alternatives as well as differential responsiveness to exogenous variables across households. The MSCL model was also statistically superior to the MNL model. In addition, the results indicate that failing to accommodate spatial correlation and unobserved response heterogeneity can lead to incorrect conclusions regarding the elasticity effects of exogenous variables.

The empirical application of the MSL model showed that the MSL structure outperforms the single-level structure both conceptually and statistically. The MSL model suggested that the social-economic and demographic composition factors generally have a smaller spatial extent of influence than the land-use factors. Between the census-unit based and the circular-unit based hierarchy, the model parameters are generally consistent in sign, but their magnitudes may differ up to 300%, indicating that, even when the multi-scale structure is used, the models are still subject to the effect of the MAUP. In the case of the effect of land-use mix on residential choice, the two spatial definitions yield contradicting and puzzling results. On the one hand, the census-unit model suggests that land-use mix has no effect on the choice behavior. On the other hand, the circular-unit model showed that land-use mix has statistically significant but opposite effect at two separate spatial scales. It remains unclear which spatial definition best explains the true behavior of decision makers.

---

is not revealed, the empirical analyses performed in this study had to assume that the characteristics of households and neighborhoods have not changed from the time the original tradeoff decision was made.

6.2 Recommendations for Further Research

As much as this study has improved the analytical methodologies for residential choice analysis, it also raises the need for further investigation on a number of research problems. These problems must ultimately be resolved if the analytical methodologies proposed in the study are to become operational for policy testing. The more significant ones of these research topics include:

1. In the empirical analyses conducted in this study, crime rates and school quality showed no significant effect on residential choice. This is most likely due to the inadequate spatial resolutions at which the two factors are specified in the models. The lack of significance of school quality may also be the result of inadequate measures of school quality and the availability of private and parochial school options (Sermons, 1998, p.122). More detailed data about these two factors should be collected and used for model estimation to prevent biased parameter estimates due to omitted variables. The inclusion of these factors at the appropriate level of detail should provide behavioral insights on how households tradeoff other choice considerations against these two important residential choice determinants.
2. Another aspect about the empirical results obtained in this study that deserves special attention is the mixed results observed about the effect of land-use variables. For the MSCL model estimated for Dallas County, Texas, all land-use variables except the percentage of land occupied by multi-family housing were statistically insignificant. For the MSL model estimated for the Bay area, contradicting results were found for the effect of land-use mix when different spatial hierarchy definitions are used. It is unclear if the difference in findings is a true reflection of the choice behavior of the observed households. Or, in the case of the MSL models, if and how the errors introduced by disaggregation and aggregation during the spatial overlay operation might have

contributed to the contradicting results. Moreover, the observed effects of land-use variables may also be the result of the competition for locations between population and employment sectors. It may appear that, for example, households are avoiding mixed land-use when in fact they are being outbid by businesses and firms for locations. Further investigation, using more detailed data and perhaps alternative analytical methods, is required to better understand the dynamics between land-use distribution and residential choice.

3. The previous two issues bring us back to the question of how to construct spatial units to appropriately capture the extent of influence of various neighborhood processes or housing market forces. The journey to finding answers to this question won't be short; and the empirical analysis performed in this study of the census-defined versus the circular units is only the starting point. With the availability of micro-level data, GIS and analytical tools such as the MSL model, future analysts studying spatial choice behavior should explore other ways of operationalizing the concept of neighborhood and investigate alternative behavioral units for spatial factors. One approach is to incorporate both concepts of fixed- and sliding-neighborhood definition in a single model. For instance, if school quality data is available, the effect of school quality is perhaps best captured over the 'fixed' (juristically defined) school catchments zones while the other factors are measured over the 'sliding' (varying by residential alternatives) circular units.
4. The two model structures proposed in this dissertation to separately address the two spatial issues arising in residential location choice models can be combined to account for spatial correlation in the disaggregate context. More specifically, by appropriately defining the paired nested structure to represent pair-wise correlations among dwelling units (as opposed to residential zones), the SCL model can be applied to the disaggregate

context where each dwelling unit is treated as a choice alternative. One can then specify the utility function using the multi-scale concept embedded in the MSL model. The resulting spatially correlated, mixed-scale, logit model will account for both the spatial correlation issue and the spatial representation issue. One can also accommodate household responsive heterogeneity by further imposing the mixing structure. However, this would greatly increase the computational burden required for model estimation.

5. The spatial structure of a metropolitan area results from a complex interaction between firms, households and institutions. This study explores only one aspect of these interactions, *i.e.* the residential location choice of households, holding other actors constant. This is a simplified approach and does not necessarily reflect the true causality (for example, Dietz (2002) found that the location of households is an important determinant of the location of employment, but the location of employment is not relevant in determining the location of households). Ways of incorporating the methodological advances introduced in this study to model more complex choice situations should be investigated.
6. The empirical analysis performed in this study is partial and static, representing a snapshot of certain urban areas at one point in time. For the residential choice models to become operational for forecasting purposes, attempts should be made to test their spatial and temporal transferability. One approach is to estimate models using data from other cities to examine whether differences in the parameter estimates across cities or time periods can be explained in terms of inherent differences in the physical or demographic structure of the cities.

BIBLIOGRAPHY

Abraham, J.E. and Hunt, J.D. (1997) Specification and estimation of a nested logit model of home, workplace and commuter mode choice by multiple worker households, *Transportation Research Record*, 1606, 17-24.

Alonso, W. (1964) *Location and Land Use*, Harvard University Press, Cambridge.

Alvanides, A., Openshaw, S. and Macgill, J. (2001) Zone design as a spatial analysis tool, Chapter 8 in *Modeling Scale in Geographical Information Science*, edited by N.J. Tate and P.M. Atkinson, John Wiley & Sons, Ltd.

Amrhein, C.G. and Flowerdew, R. (1992) The effect of data aggregation on a Poisson regression model of Canadian migration, *Environment and Planning A*, 24, 1381-1391.

Anas, A. and Chu, C. (1984) Discrete choice models and the housing price and travel to work elasticities of location demand, *Journal of Urban Economics*, 15, 107-123.

Arbia, G. (1989) Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems, Kluwer Academic, Dordrecht.

Axhausen, K. and Gärling, T. (1992) Activity-based approaches to travel analysis: conceptual frameworks, models and research problems, *Transport Reviews*, 12, 324-341.

Bach, L. (1981) The problem of aggregation and distance for analyses of accessibility and access opportunity in location-allocation models, *Environment and Planning A*, 13, 955-978.

Bagley, M.N. and Mokhtarian, P.L. (2002) The impact of residential neighborhood type on travel behavior: A structural equations modeling approach, *Annals of Regional Science*, 36, 279-297.

Banister, D. (1994) *Transport Planning*. Chapman & Hall, London.

Batty, M. and Sikdar, P.K. (1982) Spatial aggregation in gravity models: 4. Generalisations and large-scale applications, *Environment and planning A*, 14, 795-822.

Ben-Akiva, M. and Bowman, J.L. (1998) Integration of an activity-based model system and a residential location model, *Urban Studies*, 35(7), 1131-1153.

Ben-Akiva, M. and Lerman, S. (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge.

Bettman, J.R. (1979) *An Information Processing Theory of Consumer Choice*, Reading, Addison Wesley, Massachusetts.

Bhat, C.R. (2003) Random Utility-Based Discrete Choice Models for Travel Demand Analysis, *Transportation Systems Planning: Methods and Applications*, Chapter 10, 1-30, edited by K. Goulias, CRC Press.

Bhat, C.R. (2002a) Recent methodological advances relevant to activity and travel behavior, *In Perpetual Motion: Travel Behavior Research Opportunities and Application Challenges*, edited by H.S. Mahmassani, Elsevier Science, Oxford, UK, 381-414.

Bhat, C.R. (2002b) Simulation Estimation of Mixed Discrete Choice Models Using Randomized and Scrambled Halton Sequences, *Transportation Research B*, forthcoming.

Bhat, C.R. (2001) Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model, *Transportation Research Part B*, 35, 677-693.

Bhat, C.R. (1998) An analysis of travel mode and departure time choice for urban shopping trips, *Transportation Research B*, 32(6), 361-371.

Bhat, C.R. and Koppelman, F.S. (1999) A retrospective and prospective survey of time-use research, *Transportation*, 26(2), 119-139.

Blalock, H. (1964) *Causal Inferences in Nonexperimental Research*, University of North Carolina Press, Chapel Hill.

Blumen, R. (1994), Gender differences in the journey to work, *Urban Geography*, 15, 223-245.

Boarnet , M.G. and Sarmiento, S. (1998) Can land-use policy really affect travel behavior? A study of the link between non-work travel and land-use characteristics, *Urban Studies* 35(7) 1155-1169.

Boehm, T.P. (1982) A hierarchical model of housing choice, *Urban Studies*, 19, 17-31.

Bolduc, D. (1992) Generalized autoregressive errors in the multinomial probit model, *Transportation Research Part B*, 26, 155-170.

Bolduc, D., Fortin, B. and Gordon, S. (1997) Multinomial probit estimation of spatially interdependent choices: An empirical comparison of two new techniques, *International Regional Science Review*, 20, 77-101.

Börsch-Supan, A. and Hajivassiliou, V.A. (1993) Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models, *Journal of Econometrics*, 58, 347-368.

Briassoulis, H. (2000) *Analysis of Land Use Change: Theoretical and Modeling Approaches*, Regional Research Institute, West Virginia University, [http://www.rri.wvu.edu/WebBook/Briassoulis/contents.htm]

Casillas, P. (1987) Aggregation problems in location-allocation modelling, in *Spatial Analysis and Location-Allocation Models*, edited by A. Ghosh and G. Rushton, Reinhold Van Norstrand, New York, 327-344.

Chamberlain, G. (1980) Analysis of covariance with qualitative data, *Review of Economic Studies*, 47, 225-238.

Chattopadhyay S. (2000) The effectiveness of McFadden's nested logit model in valuing amenity improvement, *Regional Science and Urban Economics*, 30, 23-43.

Chu, C. (1989) A Paired combinatorial logit model for travel analysis, *Proceedings of the Fifth World Conference on Transportation Research*, 295-309, Ventura, CA.

Clark, W.A.V. and Avery, K. (1976) The effects of data aggregation in statistical analysis, *Geographical Analysis*, 8, 428-438.

Clark, W.A.V. and Onaka, J.L. (1985) An empirical test of a joint model of residential mobility and housing choice, *Environment and Planning A*, 17, 915-930.

Coulton, C.J., Korbin, J., Chan, T., and M. Su. (2001) Mapping residents' perceptions of neighborhood boundaries: a methodological note, *American Journal of Community Psychology*, 29(2), 371-383.

Crane, R. (2000) The influence of urban form on travel: An interpretive review, *Journal of Planning Literature*, 15(1), 3-23.

Daganzo, C. (1979) Multinomial Probit: The Theory and its Applications to Demand Forecasting, Academic Press, New York.

Daly, A.J. and Zachary, S. (1978) Improved multiple choice models, in *Determinants of Travel Choice*, edited by D.A. Hensher and M.Q. Dalvi, Saxon House, Westmead.

Deng, Y., Ross, S.L. and Wachter, S.M. (2003) Racial differences in homeownership: the effect of residential location, *Regional Science and Urban Economics*, 33, 517-556.

Dietz, R.D. (2002) The estimation of neighborhood effects in the social sciences: An interdisciplinary approach, *Social Science Research*, 31, 539-575.

Ding, C. (1998) The GIS-Based Human-Interactive TAZ Design Algorithm: Examining the Impacts of Data Aggregation on Transportation-Planning Analysis, *Environment and Planning B*, 25, 601-616.

Dubin, R.A. (1992) Spatial Autocorrelation and Neighborhood Quality, *Regional Science* and Urban Economics, 22, 433-452.

Earnhart, D. (2002) Combining revealed and stated data to examine housing decisions using discrete choice analysis, *Journal of Urban Economics*, 51, 143-169.

Evans, A.W. (1973) *The Economics of Residential Location*, Macmillan Press, London.

Eymann, A. and Ronning, G. (1997) Microeconometric models of tourists' destination choice, *Regional Science and Urban Economics*, 27, 735-761.

Feather, P.M. (1994) Sampling and aggregation issues in random utility model estimation, *American Journal of Agricultural Economics*, 76, 772-780.

Fischer, M. and Nijkamp, P. (1987) Spatial labour market analysis: relevance and scope, *Regional Labour Markets*, .M. Fischerand and P. Nijkamp, Elsevier Science Publishers, 1-33.

Fotheringham, A.S. (1986) Modeling hierarchical destination choice, *Environment and Planning A*, 18, 401-418.

Fotheringham, A.S. (1988) Consumer store choice and choice set definition, *Marketing Science*, **7**(3), 299-310.

Fotheringham, A.S. (1991) Statistical modeling of spatial choice: an overview, in *Spatial Analysis in Marketing: Theory, Methods, and* Applications, edited by A. Ghosh and C. Ingene, JAI Press, Greenwich, CT, 95-118.

Fotheringham, A.S., Brunsdon, C. and Charltom, M. (2000) *Quantitative Geography: Perspectives on Spatial Data Anlysis*, Sage Publications, London.

Fotheringham, A.S. and Curtis, A. (1999) Regularities in spatial information processing: implications for modeling destination choice, *Professional Geographer*, 51(2), 227-239.

Fotheringham, A.S., Densham, P.J. and Curtis, A. (1995) The zone definition problem in location-allocation modelling, *Geographical Analysis*, 27(1), 60-77.

Fotheringham, A.S. and Wong, D.W.S. (1991) The modifiable areal unit problem in multivariate statistical analysis, *Environment and Planning A*, 23, 1025-1044.

Freeman, O. and Kern, C.R. (1997) A model of workplace and residential choice in two-worker households, *Regional Science and Urban Economics*, 27, 241-260.

Gabriel, S.A. and Rosenthal, S.S. (1989) Household location and race: estimates of a multinomial logit model, *The Review of Economics and Statistics*, 17 (2), 240-249.

Galster, G.C. (2001) On the nature of neighborhood, *Urban Studies*, 38(12), 2111-2124.

Gehlke, C.E. and Biehl, K. (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material, *Journal of the American Statistical Association*, 29(Supplement), 169-170.

Goodchild, M.F. (1979) The aggregation problem in location-allocation, *Geographical Analysis*, 11(3) 240-255.

Grannis, R. (2003) Islands in the city: Social networks and street networks, working paper, Department of Sociology, University of California, Los Angeles.

Grannis, R. (1998) The importance of trivial streets: Residential streets and residential segregation, *American Journal of Sociology*, 103(6), 1530-1564.

Griliches, Z. (1961) Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change, *The Price Statistics of the Federal Government*, General Series no. 73, Columbia University Press for NBER, New York, 137-196.

Guest, A.M. and Lee, B.A. (1984) How urbanites define their neighborhoods, *Population and Environment*, 7(1), 32-56.

Guo, J.Y. (2000) A graph partitioning approach to transport analysis zone design in a geographical information system environment, Master's Thesis, Transport Research Center, Royal Melbourne Institute of Technology.

Hajivassiliou, V.A. and Ruud, P.A. (1994) Classical estimation methods for LDV models using simulations, *Handbook of Econometrics*, IV, edited by R. Engle and D. McFadden, Elsevier, New York, 2383-2441.

Hanson, S. (1986) Dimension of urban transportation problem, in *The Geography of Urban Transportation*, edited by S. Hanson, Macmillan Publishing Company, New York.

Harris, B. (1996) Land use models in transportation planning: a review of past developments and current practice, [http://www.bts.gov/other/MFD_tmip/papers/landuse/compendium/dvrpc_appb.htm]

Hansen, W.G. (1959) How Accessibility Shapes Land Use, *Journal of the American Institute of Planners*, 22(2), 73-76.

Hansen, E.R. (1987) Industrial location choice in São Paulo, Brazil, *Regional Science and Urban Economics*, 17, 89-108.

Harris, B. (1963*) Linear programming and the projection of land use*, Penn- Jersey Paper No. 20, Philadelphia, PA.

Hensher, D.A. (1999) The valuation of travel time savings for urban car drivers: Evaluating alternative model specifications, Technical paper, Institute of Transport Studies, University of Sydney.

Hensher, D.A. and Green, W.H. (2001) The mixed logit model: the state of practice and warnings for the unwary, Working paper, Institute of Transport Studies, University of Sydney.

Hillsman, E.L. and Rhoda, R. (1978) Errors in measuring distances from populations to service centers, *Annals of the Regional Science Association*, 12, 74-88.

Hirtle, S.C. and Jonides, J. (1985) Evidence of hierarchies in cognitive maps, *Memory and Cognition*, 13, 208-217.

Horowitz, J.H. (1995) Example: Modeling choices of residential location and mode of travel to work, In *The Geography of Urban Transportation*, edited by S. Hanson, Guilford Press, New York, 219-239.

Horowitz, J.H. (1991) Reconsidering the multinomial probit model, *Transportation Research B*, 25, 433-438.

Horowitz, J.H. (1981) Identification and diagnosis of specification erroe in the multinomial logit model, *Transportation Research B*, 15, 345-360.

Horton, F.E. and Reynolds, D.R. (1971) Effects of urban spatial structure on individual behavior, *Economic Geography*, 47(1), 36-48.

Hoover, E.M. and Vernon, R. (1959) *Anatomy of a Metropolis*, Harvard University Press, Cambridge, Massachusetts.

Hunt, J.D., McMillan, J.D.P. and Abraham, J.E. (1994) Stated preference investigation of influences on attractiveness of residential locations, *Transportation Research Record*, 1466, 79-87.

Jones, P. (1990) (ed.) Developments in Dynamic and Activity-Based Approaches to Travel Analysis, Avebury, Aldershot.

Jones, P., Clarke, M. and Dix, M. (1983) *Understanding Travel Behaviour*, Gower, Aldershot.

Kain, J. (1962) The journey-to-work as a determinant of residential location, *Journal of Urban Economics*, 51, 143-169.

Kain, J. and Quigley, J. (1975) *Housing Market and Racial Discrimination: A Microeconomic Analysis*, National Bureau of Economic Research, New York.

Kanaroglou, P.S. and Ferguson, M.R. (1998) The aggregated spatial choice model vs. the multinomial logit: an empirical comparison using migration microdata, *The Canadian Geographer*, 42(3), 218-231.

Kitamura, R., Mokhtarian, P.L. and Laidet, L. (1997) A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area, *Transportation*, 24, 125-158

Koppelman, F.S. and Wen, C.-H. (2000) The paired combinatorial logit model: properties, estimation and application, *Transportation Research B*, 34(2), 75-89.

Krizek, K.J. and Waddell, P. (2002) Analysis of lifestyle choices: Neighborhood type, travel patterns, and activity participation, *Transportation Research Record*, 1807, 119-128.

Lee, L.-F. (1992) On the efficiency of methods of simulated moments and maximum simulated likelihood estimation of discrete response models, *Econometric Theory*, 8, 518-552.

Lee, B.A., Campbell, K.E. and Miller, O. (1991) Racial differences in urban neighboring, *Sociological Forum*, 6(3), 525-550.

Lerman, S.R. (1983) Random utility models of spatial choice, in *Optimization and Discrete Choice in Urban Systems*, edited by B.G. Hutchinson, P. Nijkamp and M. Batty.

Lerman, S.R. (1975) A disaggregate behavioral model of urban mobility decisions, Ph.D. dissertation, Massachusetts Institute of Technology.

Levine, J. (1998) Rethinking accessibility and jobs-housing balance, *Journal of American Planning Association*, 64(2), 133-149.

Lowry, I. (1964) *A Model of Metropolis*, RM-4035-RC, The Rand Corporation, Santa Monica, CA.

Luce, R. (1959) *Individual Choice Behavor: A Theoretical Analysis*, Wiley, New York.

Lund, H. (2003) Testing the claims of New Urbanism, *APA Journal*, 69(4), 414-429.

MacDonald, H.I. (1999) Women's employment and commuting: explaining the links, *Journal of Planning Literature*, 13(3), 267–283.

Manski, C.F. (1975) Maximum score estimation of the stochastic utility model of choice, *Journal of Econometrics*, 3, 205-228.

Marschak, J. (1960) Binary choice constraints on random utility indications, in *Stanford Symposium on Mathematical Methods in the Social Sciences*, edited by K. Arrow, Stanford University Press, Stanford, CA, 312–329.

McFadden, D. (1978) Modeling the choice of residential location, *Spatial Interaction Theory and Planning Models*, edited by A. Karlqvist *et al*, North Holland Publishers, Amsterdam.

McFadden, D. (1974) Conditional logit analysis of qualitative choice behavior, *Frontiers in Econometrics*, edited by P. Zarembka, Academic Press, New York, 105-142.

McFadden, D. and Train, K. (2000) Mixed MNL models for discrete response, *Journal of Applied Econometrics*, 15(5), 447–470.

McLafferty, S. and Preston, V. (1992) Spatial mismatch and labor marker segmentation for African American and Latina women, *Economic Geography*, 68(4), 406–431.

McNamara, T.P. (1992) Spatial representation, *Geoforum*, 23, 139-150.

Miller, H.J. (1999) Potential contributions of spatial analysis to geographic information systems for transportation, *Geographical* Analysis, 20, 153-182.

Mills, E.S. (1972) *Urban Economics*, Scott Foresman, Glenview, IL.

Mitchell, R. and Rapkin, C. (1954) *Urban Traffic – A Function of Land Use*, Columbia University Press, New York.

Muth, R.F. (1969) *Cities and Housing*, University of Chicago Press, Chicago.

Nechyba, T.J. and Strauss, R.P. (1998) Community choice and local public services: A discrete choice approach, *Regional Science and Urban Economics*, 28, 51-73.

Nerella, S., and Bhat, C.R. (2003) A numerical analysis of the effect of sampling of alternatives in discrete choice models, Technical paper, Department of Civil Engineering, The University of Texas at Austin.

O'Campo, P. (2003) Invited commentary: Advancing theory and methods for multilevel models of residential neighborhoods and health, *American Journal of Epidemiology*, 157(1), 9-13.

Openshaw, S. (1996) Developing GIS-relevant zone-based spatial analysis methods, Chapter 4 of *Spatial Analysis: Modelling in a GIS Environment*, edited by P. Longley, and M. Batty, GeoInformation International, Cambridge, 55-78.

Openshaw, S. (1984) *Concepts and Techniques in Modern Geography: Number 38 - The Modifiable Areal Unit Problem*, Geo Books, Norwick.

Openshaw, S. (1977) Optimal zoning systems for spatial interaction models, *Environment and Planning A*, 9, 169-184.

Park, R. (1915) The city: Suggestions for the investigations of human behavior in the urban environment, *American Journal of Sociology*, 20(5), 577-612.

Parsons, G.R. and Hauber, A.B. (1998) Spatial boundaries and choice set definition in a random utility model of recreation demand, *Land Economics*, 74(1), 32-48.

Pellegrini, P.A. and Fotheringham A.S. (2002) Modeling spatial choice: a review and synthesis in a migration context, *Progress in Human Geography*, 26(4), 487-510.

Pozsgay, M.A. and Bhat, C.R. (2002) Destination choice modeling for home-based recreational trips: analysis and implications for land-use, transportation, and air quality planning, *Transportation Research Record* 1777, 47-54.

Price, S. (2002) Surface interpolation of apartment rental data: Can surfaces replace neighborhood mapping?, *Appraisal Journal*, July, 260-273.

Putman, S.H. and Chung, S.H. (1989) Effects of spatial system design on spatial interaction models, 1: the spatial definition problem, *Environment and Planning A*, 21, 27-46.

Quigley, J.M. (1985) Consumer choice of dwelling, neighborhood and public services, *Regional Science and Urban Economics*, 15, 41-63.

Quigley, J.M. (1976) Housing demand in the short run: An analysis of polytomous choice, *Explorations in Economic Research*, 3(1), 76-102.

Rapaport, C. (1997) Housing demand and community choice: an empirical analysis, *Journal of Urban Economics*, 42, 243-260.

Romanos, M.C. (1976) *Residential Spatial Structure*, Lexington, Lexington Books, Massachusetts.

Rosen, S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition, *Journal of Political Economy*, 82, 34-55.

Rust, R.T. and Donthu, N. (1995) Capturing geographically localized misspecification error in retail store choice models, *Journal of Marketing Research*, 32, 103-110.

Sampson, R.J., Morenoff, J.D. and Gannon-Rowley T. (2002) Assessing 'neighborhood effects': Social processes and new directions in research, *Annual Reviews in Sociology*, 28, 443-478.

Sayer, A. (1976) A critique of urban and regional modeling, *Progress in Planning,* 6(3), 191-254.

Sermons, M.W. (2000) Influence of race on household residential utility, *Geographical Analysis*, 32**(**3), 225–246.

Sermons, M.W. and Koppelman, F.S. (2001) Representing the differences between female and male commute behavior in residential location choice models, *Journal of Transport Geography*, 9, 101-110.

Sermons, M.W. and Seredich, N. (2001) Assessing traveler responsiveness to land and location based accessibility and mobility solutions, *Transportation Research D*, 6, 417-428.

Shinn, M. and Toohey S.M. (2003) Community contexts of human welfare, *Annual Reviews in Psychology*, 54, 427-459.

Small, K. (1987) A discrete choice model for ordered alternatives, *Econometrica*, 55(2), 409-424.

Srinivasan, S. and Seredich, N. (2001) Assessing traveler responsiveness to land and location based accessibility and mobility solutions, *Transportation Research D*, 6, 417-428.

Steel, D.G., Holt, D. and Tranmer, M. (1994) Modelling and Adjusting Aggregation Effects. *Proceedings of the U.S. Bureau of the Census Annual Research Conference*, U.S. Department of Commerce, Washington D.C., 382-408.

Suttles, G.D. (1972) *The Social Construction of Communities*. University of Chicago Press, Chicago.

Swait J. (2001) Choice set generation within the generalized extreme value family of discrete choice models, *Transportation Research B*, 35(7), 643-666.

Texas Education Agency (2000) *2000 Accountability Manual*.

Thill, J.C. (1992) Choice set formation for destination choice modeling, *Progress in Human Geography*, 16, 361-382.

Thill, J.-C. and Wheeler, A. (2000) Tree induction of spatial choice behavior, *Transportation Research Record*, 1719, 250-258.

Thurstone, L. (1927) A law of comparative judgment, *Psychological Review*, 34, 273–286.

Tobler, W. (1991) Frame Independent Spatial Analysis, in *Accuracy of Spatial Databases*, edited by M. Goodchild, and S. Gopal, Taylor and Francis, New York, 115-122.

Tobler, W. (1970) A Computer Model Simulating Urban Growth in the Detroit Region, *Economic Geography*, 46(2), 234-240.

Train, K., 2003, *Discrete Choice Methods with Simulation*, Cambridge University Press.

Train, K. (1999) Halton sequences for mixed logit, technical paper, Department of Economics, University of California, Berkeley.

Train, K. (1998) Recreational demand models with taste differences over people, *Land economics*, 74(2), 230-239.

Tu, Y. and Goldfinch, J. (1996) A two-stage housing choice forecasting model, *Urban Studies*, 33(3), 517-537.

Turner, T. and Niemeier, D. (1997) Travel to work and household responsibility: new evidence, *Transportation*, 24, 397–419.

Vovsha, P. (1997) The cross-nested logit model: application to ode choice in the Tel-Aviv metropolitan area, *Transportation Research Record*, 1607, 6-15.

Waddell, P. (2000) Towards a behavioral integration of land use and transportation modeling, presented at the *9th International Association for Travel Behavior Research Conference*, Queensland, Australia.

Waddell, P. (1996) Accessibility and residential location: the interaction of workplace, residential mobility, tenure, and location choices, presented at the *Lincoln Land Institute TRED Conference*. [http://www.odot.state.or.us/tddtpan/modeling.html]

Waddell, P. (1993) Exogenous workplace choice in residential location models: Is the assumption valid?, *Geographical Analysis*, 25, 65-82.

Waddell, P. (1992) A multinomial logit model of race and urban structure, *Urban Geography*, 13(2), 127-141.

Weaton, W.C. (1974) Linear programming and locational equilibrium: the Herbert-Stevens model revisited, *Journal of Urban Economics*, 1, 278-287.

Weisbrod, G., Lerman, S.R. and Ben-Akiva, M. (1980) Tradeoffs in residential location decisions: Transportation versus other factors, *Transport policy and Decision Making*, 1, 13-26.

Wen, C.-H. and Koppelman, F.S. (2001) The generalized nested logit model**,** *Transportation Research B*, 35(7) 627-641.

White, M.J. (1977) A model of residential location choice and commuting by men and women workers, *Journal of Regional Science*, 17(1), 41-52.

Williams, H.C.W.L. (1977) On the formation of travel demand models and economic evaluation measures of user benefit, *Environment and Planning*, 9A, 285-344.

Wong, D. (1996) Aggregation effects in geo-referenced data', Chapter 5 of *Practical Handbook of Spatial Statistics*, edited by S. Arlinghaus, *et al.*, CRC Press, Boca Raton, Florida.

Xue, Y. and Brown, D.E. (2003) A decision model for spatial site selection by criminals: a foundation for law enforcement decision support, *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Application and reviews*, 33(1), 78-85.

Yule, G.U. and Kendall, M.G. (1950) A*n Introduction to the Theory of Statistics*, Griffin, London.

VITA

Jessica Yingchieh Guo was born in Taipei, Taiwan on March 27, 1972, the daughter of Mark Lee-Jer and Shirley Hsiulan Guo. She migrated with her family to Australia in 1989. After completing her secondary education at Caulfield Grammar School, Melbourne, she entered Melbourne University in 1991. She received the degree of Bachelor of Science with Honors from Melbourne University in December 1995. During the following four years, she studied on a part-time basis at the Transport Research Center, Royal Melbourne Institute of Technology. She was employed as an experimental scientist for the Division of Building, Construction and Engineering, CSIRO, Australia, from 1997 to 2000. She also worked on a part-time basis as a tutor during 1998 and 1999 at Melbourne University and Swinburne University of Technology. In February 2000, she received the degree of Master of Business (by research) and the University Research Prize from Royal Melbourne Institute of Technology. She then worked briefly as a research fellow at the Transport Research Center, Royal Melbourne Institute of Technology, before entering the Graduate School of the University of Texas in September 2000.

Permanent Address: 1719 Coles Farm Drive, Sugar Land, Texas 77478

This dissertation was typed by the author.